Evolution of Thoughts

Kaneko, M.

Publication date:
1998

Evolution of Thoughts: Deductive Game Theories in the Inductive Game Situation Part I

Mamoru Kaneko*
Institute of Policy and Planning Sciences, University of Tsukuba,
Ibaraki 305, Japan (kaneko@shako.sk.tsukuba.ac.jp)

June, 1998

Abstract

We consider an evolution of individual thoughts on decision making in a recurrent game situation. Deductive game theories are viewed as constructed inductively by an individual player. It is a basic assumption that each player has no *a priori* belief/knowledge on the game structure but has recorded his experiences induced by trial and error. The individual player constructs a deductive theory from his experiences (and/or from some other source of knowledge such as communications with other players) so as to explain his (and/or others') observed behavior as deductively decided. This paper considers, specifically, the recurrent situation of a finite 2-person game with perfect monitoring, where each player forms inductively beliefs on his own payoff function and, maybe, on the other's. We start with the case where only individual experiences are available to a player, and go to cases where more information is available. In this context, we discuss an evolution of "deductive theories" in the mind of the individual player for his decision making, and examine the roles of deductive and inductive reasonings in the evolution process.

In Part I, we will start with Phase 0 where the player has recorded his experiences and does not construct a theory, and then will discuss Phase 1 where a player thinks as if he is a one-person decision maker.

*The research of this paper was partially supported by the Tokyo Center of Economic Research. The author thanks A. Matsui, T. Nagashima and D. DeJong for valuable discussions on the subjects of this paper.

We will consider also Phase 2 where a player makes a decision based on his prediction on the other's by regarding the other as a one-person decision maker. Then we discuss whether he may notice the need to go to further phases, which are the subjects of Part II.

1 Introduction

We consider an evolution of individual thoughts on decision making in a recurrent game situation. Deductive game theories are viewed as constructed inductively by an individual player in this recurrent situation, where "induction" means an extension of limited experiences into a general law and "deduction" means logical and mathematical reasonings from accepted belief/knowledge. Our consideration enables us to discuss the emergence and evolution of thoughts required for decision making. Also, we argue that deduction plays important but limited roles in the evolution process of thought making, and that players realize by deduction, in some cases but not always, the need of other experiential sources of information. We will also find several pitfalls where an individual thought may get stuck inductively or deductively.

Specifically, we will develop a connection between the inductive game theory given by Kaneko-Matsui [5] and the game logic approach initiated by Kaneko-Nagashima [6] and [7]. The former focuses on rule-governed behavior and inductive constructions of images on the society in the recurrent situation, while the latter treats individual *ex ante* decision making in the one-shot game situation. As in the inductive game theory, we start with the assumption that each player has no *a priori* belief/knowledge on the structure of the game including his own payoff function but has recorded his experiences induced by trial and error. Then he constructs a "theory" by hypothesizing inductively some belief/knowledge from his experiences, and deduces logical consequences from

Phase 0 (**Subsection 2.2**): **Inductive Decision Making without a Deductive**

Theory

↓

Phase 1 (**Section 5**): **One-Person Decision Making**

↓

Phase 2 (**Section 6**): **Regarding the Other Player as a One-Person Decision**

Maker

$\downarrow$

$\vdots$

Phase k (Part II): Reciprocally Interactive Decision Making: Transitory

Phase

$\downarrow$

$\vdots$

Phase ω (Part II): Reciprocally Interactive Decision Making: Limit Phase

Diagram 1.1

his theory. If he notices that such logical consequences are incompatible with his experiences, he would modify his theory. In this sense, the individual player takes the *hypothetico-deductive method* in the inductive game situation. The evolution of such individual theories we discuss takes the form of Diagram 1.1. To discuss this diagram, we first review the approaches of the inductive game theory and game logic.

For the present, the distance between the inductive game theory and game logic approach is quite large. The reason for the distance is that the former has targeted on complicated societal games with little sophisticated players, while the latter has focussed on transcendental parts of game theoretical decision making with very sophisticated players. In Diagram 1.1, the game logic approach has focussed mainly on Phase ω, while the inductive game theory has started with Phase 0 and has gone to a slightly different direction. The complicated societal game Kaneko-Matsui [5] targeted the feature that the individual player can observe only the values of macro variables in addition to micro variables in his social world almost negligible relative to the entire society. They discussed possible images on the society to be built by the individual player from his experiences, which described worlds external to his and other players' minds.[1] The game logic approach emphasizes the internal subjective and intersubjective thinkings of individual players. The game logic approach is suitable, probably, to small micro games, but not to the discussions on a player's image from the macro aggregated observations. In this paper, therefore, we restrict ourselves to consider

[1]The image is formulated as a model of such a game situation which is requied to be coherent with the individual experiences. The model does not directly describe epistemic aspects of players, i.e., the model is rather close to the stadard game theory.

the recurrent situation of, particularly, a finite 2-person game with perfect monitoring in addition to his own payoff values.

Small Micro Games ⟵ ··· ⟶ Large Societal Games

Diagram 1.2

To have a connection between the inductive game theory and game logic approach, we need to relativize transcendencies involved in the game logic approach. The first is that "knowledge", instead of "belief", is virtually used in the axiomatic consideration of decision making in the game logic approach, though the game logic itself of Kaneko-Nagashima [7] is developed so as to distinguish them explicitly. The other transcendency involved is the common knowledge assumption of the behavior principle as well as the game structure. With these, the axiomatic considerations of decision making are given in Kaneko-Nagashima [7] and Kaneko [2]. These make a great gap between the game logic approach and inductive game theory. We will fill this gap by distinguishing between "belief" and "knowledge" in our axiomatic treatment of decision making, and also by treating the common knowledge assumption as a limit case, corresponding to Phase ω in Diagram 1.1, among alternatives cases.

We distinguish between *belief and knowledge* so that knowledge is true belief and belief is required to be only logically consistent. We do not require knowledge to be verifiable by a player, and the truth is referred to the outsider's point of view.[2] This distinction enables us to describe an interpersonal feature of belief and knowledge in games that each player knows (truly believe) his own payoff function at least up to his experiences, while he cannot experience the other player's payoffs at all and construct some beliefs on it from the observables for him and/or information obtained via some communication.

The common knowledge assumption is relativized as necessary for some games and as unnecessary for some other games by Kaneko [3] and [4]. For example, if a game allows a player to have a dominant strategy, he could choose a *payoff-maximizing* strategy without predicting (thinking about) the other player's decision. This argument can be applied to either player in the game of Table 1.1 (Prisoner's Dilemma), and only to player 2 in the game of Table 1.2. In the first game, the belief/knowledge of depth

[2]Sometimes, knowledge is also required to be verifiable. The definitions of knowledge is still disputed in the literature of philosophy (cf., p. 447, [1]).

1, i.e., no interpersonal consideration, is required. In the second game, player 1 can make a decision by applying the same argument to player 2, which means that the belief/knowledge of depth 2 is required but not the common knowledge. The dominant strategy argument can be applied to neither player in the game of Table 1.3, which appears to require the common knowledge assumption.

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$
$\mathbf{s}_{11}$	$(5,5)$	$(1,6)$
$\mathbf{s}_{12}$	$(6,1)$	$(3,3)^*$

Table 1.1

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$
$\mathbf{s}_{11}$	$(5,5)$	$(1,6)$
$\mathbf{s}_{12}$	$(2,1)$	$(3,3)^*$

Table 1.2

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$
$\mathbf{s}_{11}$	$(5,5)$	$(1,2)$	$(3,4)$
$\mathbf{s}_{12}$	$(6,1)$	$(3,3)^*$	$(0,2)$

Table 1.3

The notion of an *interaction structure* introduced by Kaneko [3] describes whether or not the individual player needs to think about the other player's decision making. In the 2-person case, there are essentially three interaction structures: the first two correspond to Phases 1 and 2, and the last one corresponds to the Transitory Phase k and Limit Phase ω. These are illustrated by the above three examples. In Phase 1, the individual player follows the dominant strategy behavior principle, and in Phase 2, he hypothesizes that the other player thinks in the way of Phase 1. In Phases k and ω, the players think reciprocally about each other.[3,4]

By the distinction between belief and knowledge, the player's thinking can be treated in the subjective manner. Hence, we can regard the above

[3]Some reader may find that the arguments here are closely related to dominant-solvability (cf., Moulin [10]). Kaneko [3] made comparisons with our approach with dominant-solvability in the n-person case. Iterating elimination of dominated strategies as well as dominant-solvability are relevant in the epistemic consideration of decision making. However, since these subjects slightly deviates from our main concern, we do not consider them in this paper.

[4]In fact, it is shown in [3] and [4] that there is a great spectrum of games and prediction structures required for individual decision making. The objective of [3] and [4] was to give a unified treatment of such a great spectrum of prediction structures.

dominant and nondominant strategy arguments as undertaken by an individual player. The evolution of individual thoughts will be examined from the viewpoints of interactions between the subjective thinking and objective experiences.

With these preparations, we can start our discussions on evolutions of individual thoughts. Since the discussions are quite long, we divide the present paper into two parts. In Part I, we prepare the basic game theoretical notions, inductive game theory and game logic approach. Then we formulate a set of behavioral axioms which is classified into three types based on the three interaction structures mentioned above. In Part I, we consider the two simplest cases, Phases 1 and 2, where an individual player follows the dominant strategy behavior principle and where a player regards the other as following this principle. These considerations would not be vacuous even if each player's true payoff function allows him to have no dominant strategies. There may be gaps between the individual subjective thinking and our (outsiders') objective thinking. These cases may look quite restrictive in the objective game theoretic sense. However, our purpose is to consider to what extent an individual player can notice such gaps and needs to go to further phases.

An individual player *may* go from Phase 0 to 1 or 2 and/or from 1 to 2 without much difficulty. However, there may be still a possibility for him *to get stuck in the pitfall of developing false beliefs* due to the lack of information. The player himself may (at least potentially) deductively notice this pitfall, and if so, he may notice that deduction is insufficient and he needs more experiential source for knowledge.

In Part II, we will discuss the Transitory Phase k and Limit Phase ω. After leaving Phase 2, a lot of difficulties both inductive and deductive are waiting for a player. He may and/or may not notice that he gets involved in a difficulty. The Limit Phase ω is difficult to reach but once a player reaches this phase, some difficulties are resolved. However, it might be an illusion for him that all the problems are resolved. For example, it will be shown in Part II that both players believe the common knowledge of the game structure, where the contents of their common knowledge are different from each other and actually are false from the objective point of view. Also, even if these problems are solved, more difficult problems may be waiting for him, which are transcendental problems discussed in Kaneko-Nagashima [7] and Kaneko [2]. In Part II, we discuss our evolution process before such transcendental problems.

2 Game Theoretical Concepts and Inductive Game Theory

In Subsection 2.1 we give basic game theoretical concepts, and then briefly review the inductive game theory of Kaneko-Matsui [5] in Subsection 2.2.

2.1 Basic Game Theoretical Concepts

Consider a 2-person finite noncooperative game g in strategic form. The players are denoted by $1, 2$, and each player i has ℓ_i pure strategies. We assume that the players do not play mixed strategies. Player i's strategy space is denoted by $\Sigma_i := \{\mathbf{s}_{i1}, ..., \mathbf{s}_{i\ell_i}\}$. His *payoff function* is a rational number valued function g_i on $\Sigma_i \times \Sigma_j$ for $i, j = 1, 2$ $(i \neq j)$, that is, g_1 is defined on $\Sigma_1 \times \Sigma_2$ and g_2 on $\Sigma_2 \times \Sigma_1$. Similar, slightly unusual, notations will be used in several places. We call a vector $a = (a_1, a_2)$ in $\Sigma = \Sigma_1 \times \Sigma_2$ a *strategy profile*. By $g_i(a)$, we denote $g_i(a_i; a_j)$.

Let $(a_1, a_2) \in \Sigma$ and $i, j = 1, 2$ $(i \neq j)$. We say that a_i is a *best response to* a_j iff

$$g_i(a_i; a_j) \geq g_i(b_i; a_j) \text{ for all } b_i \in \Sigma_i.$$

We say that a_i is a *dominant strategy* iff a_i is a best response to a_j for all $a_j \in \Sigma_i$. A strategy profile $a = (a_1, a_2)$ is called a *Nash equilibrium* iff each a_i is a best response to the other a_j.

In the game of Table 1.1, $\mathbf{s}_{i2}$ is a dominant strategy for $i = 1, 2$. In the game of Table 1.2, $\mathbf{s}_{12}$ is a dominant strategy, and $\mathbf{s}_{22}$ is a best response to $\mathbf{s}_{12}$ but is not a dominant strategy. In both games, $(\mathbf{s}_{12}, \mathbf{s}_{22})$ is a unique Nash equilibrium. In the game of Table 1.3, neither player has a dominant strategy, but $(\mathbf{s}_{12}, \mathbf{s}_{22})$ is still a unique Nash equilibrium.

As discussed in Section 1, a player may or may not need to predict the other player's decision making. To discuss this interpersonal feature of predictions, we consider the following concept. An *interaction structure for player* i is a vector $\mathcal{J} = (J_1, J_2)$ satisfying

$$j \in J_j \subseteq \{1, 2\} \text{ for } j = 1, 2. \tag{1}$$

This means that if $j \in J_i$, then i needs to predict j's decision and, if $i \in J_j$ in addition to $j \in J_i$, then i believes that j needs to predicts i's decision. This notion is introduced by Kaneko [3] for general n-person games and is considered from the game theoretical point of view. In the 2-person case,

there are the following four interaction structures:

$$\mathcal{J}^1 = (\{1\},\{2\}),\ \mathcal{J}^2 = (\{1\},\{1,2\}),$$
$$\mathcal{J}^3 = (\{1,2\},\{2\}),\ \text{and}\ \mathcal{J}^4 = (\{1,2\},\{1,2\}). \tag{2}$$

In this paper, we attribute an interaction structure to a single player, though the formalism is free from this subjective interpretation. From this subjective point of view, the first two are equivalent for player 1, and the first and third are equivalent for player 2. We keep this redundancy for presentational purpose. From player 1's point of view, these interaction structures $\mathcal{J}^1$, $\mathcal{J}^3$ and $\mathcal{J}^4$ mean

(1): player 1 believes that he can make a decision without thinking about 2's decision making;

(2): player 1 believes that he needs to predict 2's decision and believes that 2's belief on decision making takes the form (1);

(3): player 1 believes that he needs to predict 2's decision and believes that 2's belief on decision making takes the reciprocal form.

If we adopt the dominant strategy behavior principle in the Prisoner's Dilemma, then (1) is applied to both players. When $\mathcal{J}^1 = (\{1\},\{1,2\})$ is applied to (1), the second element is irrelevant for player 1. In the game of Table 1.2, player 1 needs to predict 2's decision making and 2 can make a decision without predicting 1's decision. This case is expressed with $\mathcal{J}^3 = (\{1,2\},\{2\})$. The last interaction structure $\mathcal{J}^4 = (\{1,2\},\{1,2\})$ corresponds to the case that either player has no dominant strategies and needs to predict the other's decision such as in the game of Table 1.3.

2.2 Brief Review of Inductive Game Theory

Here we give a brief description of the inductive game theory of Kaneko-Matsui [5]. In the recurrent situation of the game $g = (g_1, g_2)$,

unilateral trials

past $\quad \cdots \quad g \quad \cdots \quad g \quad \cdots \quad g \quad \cdots$ future

we consider a stationary state (strategy profile) $a^* = (a_1^*, a_2^*)$, subject to unilateral deviations of an individual player. Kaneko-Matsui assume the following postulates, which are slightly modified for our situation.

Postulate 1: After each play of game g, player i observes the outcome (a_1, a_2) and his payoff value $g_i(a)$ if the players chose $a = (a_1, a_2)$.

Postulate 2: Both players know that each player i has ℓ_i strategies. Each i *has* the payoff function $g_i(\cdot)$, but has no *a priori* knowledge on it.

Postulate 3: Each player i behaves according to his strategy a_i^*, subject to (stochastic) trial deviations with small probabilities once in a while, but after each trial, he returns to his stationary strategy a_i^* (unless his experiences tell that it might be better to deviate).[5]

Postulate 4: Each player records the experiences induced by his and other player's trials.

Now we assume the following postulate and will relax it in Section 6 and Part II.

Postulate 5: Events of trials simultaneously made by the two players have negligible frequencies and are ignored by the players.

Under these postulates, the *individual experiences* in the past are formulated as

$$\mathcal{E}(i \mid a^*) = \left\{ [(a_i; a_j), g_i(a_i; a_j)] : a_i = a_i^* \text{ or } a_j = a_j^* \right\}.$$

These consist of three parts:

(a): the *stationary* experience: $[(a_i^*; a_j^*), g_i(a_i^*; a_j^*)]$ is usually experienced by player i;

(b): the *active* experiences: $\{[(a_i; a_j^*), g_i(a_i; a_j^*)] : a_i \neq a_i^*\}$ is the set of experiences induced by player i's own deviations;

(c): the *passive* experiences: $\{[(a_i^*; a_j), g_i(a_i^*; a_j)] : a_j \neq a_j^*\}$ is the set of experiences induced passively by j's deviations.

Postulate 2 states that each player has no *a priori* knowledge on his own payoff function, though he has it. Nevertheless, he has experienced various payoff values in $\mathcal{E}(i \mid a^*)$, and can associate the observed strategy pairs with the received payoffs. In this sense, he knows his own payoff function up to the experienced domain. This is still obtained by induction. By this knowledge, he may maximize his payoffs: if he has found a higher payoff value which can be induced by his own trial, he has an incentive to deviate

[5] Here we exclude the possibility that the behavior pattern of a player suggests to react to a deviation of the other player, for example, Markov behavior is excluded. In Kaneko-Matsui [5], such a possiblity is taken into account by formulating a situation in terms of an extensive game.

from his present stationary behavior a_i^*. Therefore, we make a postulate on his behavior in such a case, which defines the stability of $a^* = (a_1^*, a_2^*)$.

Postulate 6.(1): If no active experience gives a higher payoff to player i than his stationary payoff $g_i(a^*)$, then he continues playing a_i^* (still subject to his occasional trials).

(2): If some active experience $[(a_i; a_j^*), g_i(a_i; a_j^*)]$ gives a higher payoff to player i than his stationary payoff $g_i(a^*)$, then he would increase intentionally (maybe, slightly or drastically) the frequency of the deviation a_i.

The following definition is based on this postulate. We say that player i has an *incentive for an intentional deviation* in a^* iff there is an active experience $[(a_i; a_j^*), g_i(a_i; a_j^*)]$ with $g_i(a_i, a_j^*) > g_i(a^*)$. A strategy profile a^* is an *inductively stable state* iff no player has an incentive for an intentional deviation. Kaneko-Matsui [5] observed the following simple fact.

Proposition 2.1. The stationary state $a^* = (a_1^*, a_2^*)$ is inductively stable if and only if it is a Nash equilibrium in game $g = (g_1, g_2)$.

Thus, inductive stability is simply a translation of the mathematical definition of Nash equilibrium.

Now our question is how individual player i constructs a theory on the game situation $g = (g_1, g_2)$ he is playing. According to Postulate 2, he can infer that the set of possible outcomes in this game situation is given as $\Sigma = \Sigma_1 \times \Sigma_2$, and consequently that he would receive a payoff value in each outcome in Σ. Therefore we assume that player i considers a *total* payoff function $\tilde{g}_i$ for him. That is, $\tilde{g}_i$ can be a candidate for a payoff function in player i's thought when only his experiences $\mathcal{E}(i \mid a^*)$ are available to him. Then he may explain his behavior by this believed payoff function. This will be explicitly discussed in Section 5.3.

An individual player is going to construct a "theory" on the game, and it may include the description of the behavior of the other player. For his own and the other player's behavior, we assume the following postulate.

Postulate 7. An individual player constructs a theory to explain his own behavior and, maybe, the observed behavior of the other player. His theory has not only a description of the external world but also descriptions of his and the others' internal worlds (minds) as well. He also assumes that he himself is a payoff maximizer and that if his theory includes the other player's behavior, the other player is also a payoff maximizer. If his theory includes the other player's prediction on himself, then the same (or symmetric in a general sense) principle is assumed.

(2): The payoff maximizing requirement should be consistent with his observations, i.e., the derived recommendation from the theory suggests the stationary strategy a_i^* (and a_j^* for the other player).

It is the feature peculiar to small micro societies to introspect about internal worlds in the players' minds. One reason is that a face-to-face communication is possible in a small micro society. Such an introspection may be impossible and is probably irrelevant in a large societal situation. This differentiates the Kaneko-Matsui [5] approach from the present approach.[6]

Throughout the present paper, we assume that the game $g = (g_1, g_2)$ has a Nash equilibrium in pure strategies. In the end of Part II, we will give a remark on the game without a Nash equilibrium in pure strategies.

3 Game Logic Approach

In this section, we prepare the formalized language $\mathcal{P}$ in which the game theoretical notions given in Section 2 can be expressed, and then present propositional epistemic logic KD4^2. In Subsection 3.2, we redescribe game theoretical concepts in $\mathcal{P}$. In the last subsection, we give a definition of the $\mathcal{J}$-belief set for a player.

3.1 Epistemic Logic KD4^2

In the following, we use some notions of predicate logic. Since, however, we use neither variables nor quantifiers, the following logic is essentially propositional.

We start with the following list of symbols:

constant symbols: $\mathbf{s}_{11}, ..., \mathbf{s}_{1\ell_1}$; $\mathbf{s}_{21}, ..., \mathbf{s}_{2\ell_2}$;

4-*ary predicate symbols*: R_1, R_2 ;

unary predicate symbols: $I_{11}, I_{12}, I_{21}, I_{22}$;

belief operator symbols: B_1,B_2 ;

logical connectives: $\neg$ (not), $\supset$ (implies), $\bigwedge$ (and), $\bigvee$ (or) ;

parentheses: (,).

[6] Kaneko-Matsui [5] considered an individual model on the society without including the other players' utility functions. This does not demarcate between their approach and that of the present paper. The demarcation is made by considering internal worlds in minds or not.

As in Subsection 2.1, the constants $\mathbf{s}_{11}, ..., \mathbf{s}_{1\ell_1}$, $\mathbf{s}_{21}, ..., \mathbf{s}_{2\ell_2}$ are the players' pure strategies. The 4-ary symbol $R_i(\cdot : \cdot)$ is used to express player i' payoff function g_i. The unary symbol $I_{ij}(\cdot)$ is to describe i's prediction of player j' strategy choice, that is, $I_{ij}(a_j)$ means that i predicts that j could choose a_j as a final decision in his decision making. Of course, a_i itself is i's own possible decision. These $I_{11}, I_{12}, I_{21}, I_{22}$ will be determined by the nonlogical axioms which will be given in Section 4. By the expression $\mathrm{B}_i(A)$, we mean that player i believes formula A.

First, we develop the space of formulae. For strategy profiles $a = (a_1, a_2), b = (b_1, b_2)$ in Σ, the expressions $R_i(a_i, a_j : b_i, b_j)$ $(\{i, j\} = \{1, 2\})$ and $I_{ij}(a_j)$ $(i, j \in \{1, 2\})$ are *atomic formulae*. These atomic formulae correspond to propositional variables in the standard formulation of propositional logic. Since the number of strategies is finite, so is the number of atomic formulae. Similar to the previous notation, we denote $R_i(a_i, a_j : b_i, b_j)$ by $R_i(a : b)$.

Let $\mathcal{P}$ be the set of all formulae generated by the standard finitary inductive definition with respect to $\neg, \supset, \bigwedge, \bigvee$ and B_1,B_2 from the atomic formulae. That is, $\mathcal{P}$ is the set of *formulae* defined by the following induction:

(**0-i**): any atomic formula is a formula;

(**0-ii**): if A and B are formulae, so are $(\neg A), (A \supset B)$ and $\mathrm{B}_i(A)$, and if Φ is a finite nonempty set of formulae, then $(\bigwedge \Phi)$ and $(\bigvee \Phi)$ are also formulae.[7]

We abbreviate $\bigwedge\{A, B\}$ and $\bigvee\{A, B\}$ as $A \bigwedge B$ and $A \bigvee B$, and $(A \supset B) \bigwedge (B \supset A)$ as $A \equiv B$, etc. We also abbreviate some parentheses in the standard manner.

Base logic $\mathrm{GL_0}$ is defined by the following five axiom schemata and three inference rules: for any formulae A, B, C and finite nonempty set Φ of formulae,

(L1): $A \supset (B \supset A)$;

(L2): $(A \supset (B \supset C)) \supset ((A \supset B) \supset (A \supset C))$;

(L3): $(\neg A \supset \neg B) \supset ((\neg A \supset B) \supset A)$;

(L4): $\bigwedge \Phi \supset A$, where $A \in \Phi$;

(L5): $A \supset \bigvee \Phi$, where $A \in \Phi$;

[7] and every formula is obtained by a finite number of applications of these steps. We will not add this qualification in the following inductive definitions.

$$\frac{A \supset B \quad A}{B}\ (\mathrm{MP})$$

$$\frac{\{A \supset B : B \in \Phi\}}{A \supset \bigwedge \Phi}\ (\bigwedge\text{-Rule}) \qquad \frac{\{A \supset B : A \in \Phi\}}{\bigvee \Phi \supset B}\ (\bigvee\text{-Rule}).$$

These axioms and inference rules determine base logic $\mathrm{GL_0}$, which is actually classical propositional logic.

We define epistemic logic $\mathrm{KD4}^2$ by adding the following axiom schemata and inference rule to $\mathrm{GL_0}$: for any formulae A, B and $i = 1, 2$;

(MP_i): $\mathrm{B}_i(A \supset C) \bigwedge \mathrm{B}_i(A) \supset \mathrm{B}_i(C)$;

(D_i): $\neg \mathrm{B}_i(\neg A \bigwedge A)$;

(PI_i): $\mathrm{B}_i(A) \supset \mathrm{B}_i\mathrm{B}_i(A)$;

(Necessitation): $\dfrac{A}{\mathrm{B}_i(A)}$.

We will abbreviate Necessitation as Nec, and use MP_i, D_i, PI_i as generic names for these with different $i = 1, 2$.

A *proof* P in $\mathrm{KD4}^2$ is a finite tree with the following properties: (i) a formula is associated with each node, and the formula associated with each leaf is an instance of the above axioms; and (ii) adjoining nodes together with their associated formulae form an instance of the above inferences. We write $\vdash A$ iff there is a proof P such that A is associated with the root of P. *Nonlogical axioms* (e.g., game theoretical and/or some mathematical axioms) are introduced as follows. For any subset Γ of $\mathcal{P}$, we write $\Gamma \vdash A$ iff $\vdash \bigwedge \Phi \supset A$ for some nonempty finite subset Φ of Γ.[8] When Γ is empty, $\Gamma \vdash A$ is assumed to be $\vdash A$ itself. We also abbreviate $\Gamma \cup \Theta \vdash A$ and $\Gamma \cup \{B\} \vdash A$ as $\Gamma, \Theta \vdash A$ and $\Gamma, B \vdash A$, etc.

Axiom MP_i and inference rule Nec in addition to $\mathrm{GL_0}$ give the complete logical ability to each player (see [7]). Axiom D_i requires that each player's beliefs are consistent, which is further discussed below. Axiom PI_i, called the *Positive Introspection*, means that if player i knows A, he knows that he knows A.

[8]Since $\mathrm{KD4}^2$ has Nec, nonlogical axioms should be introduced in this manner, instead of being initial formulae in a proof. For the treatment of nonlogical axioms in a logic with Nec, see Kaneko-Nagashima [9].

Necessitation looks quite demanding, since it implies, without restricting its use, that every provable formula in $\mathrm{KD4}^2$ is, virtually, common knowledge. This bare use of Nec is not very suitable in our evolutionary context. However, we can impose restrictions on the use of Necessitation in each evolutionary phase: Necessitation is used with formulae with more repetitions of belief operators as evolution gets progressed. This will be discussed in Section 3 of Part II.

We list the basic facts and will use them without references (see Kaneko-Nagashima [7]).

Lemma 3.1. Let Γ, Θ be sets of formulae, Φ a finite set of formulae, and A, B, C formulae. Then

(1): if $\Gamma \vdash A \supset B$ and $\Theta \vdash B \supset C$, then $\Gamma, \Theta \vdash A \supset C$;

(2): $\vdash (A \bigwedge B \supset C) \equiv (A \supset (B \supset C))$;

(3): $\vdash \bigwedge \Phi$ if and only if $\vdash A$ for all $A \in \Phi$;

(4): $\vdash \mathrm{B}_i(\bigwedge \Phi) \equiv \bigwedge \mathrm{B}_i(\Phi)$; where $\mathrm{B}_i(\Phi)$ is the set $\{\mathrm{B}_i(A) : A \in \Phi\}$;

(5): $\vdash \bigvee \mathrm{B}_i(\Phi) \supset \mathrm{B}_i(\bigvee \Phi)$;

(6): $\vdash \mathrm{B}_i(\neg A) \supset \neg \mathrm{B}_i(A)$.

It is important to remark that in our logic, we do not assume the following axiom schema:

(T_i): $\mathrm{B}_i(A) \supset A$.

which is called the *Veridicality* (or *Truthfulness*) *Axiom*. When we add this axiom to $\mathrm{KD4}^2$, the resulting logic is called epistemic logic $\mathrm{S4}^2$, where beliefs are true from the reference viewpoint. In this sense, $\mathrm{B}_i(A)$ means that player i knows A, and we cannot distinguish beliefs from knowledge in $\mathrm{S4}^2$. On the contrary, $\mathrm{KD4}^2$ allows us to capture the concept of knowledge discussed in $\mathrm{S4}^2$ as well as beliefs. More precisely, we denote the following formula by $\mathrm{B}_i^+(A)$:

$$\mathrm{B}_i(A) \bigwedge A. \tag{3}$$

Then this new operator $\mathrm{B}_i^+(\cdot)$ captures epistemic logic $\mathrm{S4}^2$ in the sense that it satisfies the epistemic axioms, Necessitation as well as the axiom T_i. In fact, it is also proved that epistemic logic $\mathrm{S4}^2$ with knowledge operator symbols K_1 and K_2 can be embedded into our $\mathrm{KD4}^2$ by translating $K_i(A)$ into $\mathrm{B}_i(A^*) \bigwedge A^*$ (A^* is obtained by the translation of the same principle).

For our game theoretical purpose, we would like to distinguish between belief and knowledge – true belief. For example, the following two formulae

$$\mathrm{B}_i^+(g_i) \text{ and } \mathrm{B}_i\mathrm{B}_j^+(g_j)$$

mean player i's knowledge on his own payoff function and i's belief on j's knowledge on j's payoff function g_j. These are treated as nonlogical axioms. Here g_i and g_j as formulae will be introduced in Subsection 3.2.

The following will be used.

Lemma 3.2. Let Γ be a set of formulae and A a formula. Then

(1): if $\Gamma \vdash A$, then $\mathrm{B}_i(\Gamma) \vdash \mathrm{B}_i(A)$, where $\mathrm{B}_i(\Gamma) = \{\mathrm{B}_i(C) : C \in \Gamma\}$;

(2): $\vdash \mathrm{B}_i^+(A) \supset \mathrm{B}_i^+\mathrm{B}_i^+(A)$;

(3): $\mathrm{B}_i^+(\Gamma) \vdash A$ implies $\mathrm{B}_i^+(\Gamma) \vdash \mathrm{B}_i(A)$ (and a fortiori, $\mathrm{B}_i^+(\Gamma) \vdash \mathrm{B}_i^+(A)$).

For later purposes, we mention the completeness-soundness result for the nonepistemic fragment. We say that a formula A is *nonepistemic* iff neither B_1 nor B_2 occurs in A. We denote the set of all nonepistemic formulae by $\mathcal{P}^N$. By restricting the axioms and inference rules to those of base logic $\mathrm{GL}_\mathbf{0}$ in $\mathcal{P}^N$, we define the provability relation of $\mathrm{GL}_\mathbf{0}$, which is denoted by $\vdash_\mathbf{0}$. This is classical logic, and is sound and complete with respect to two-valued semantics

An *assignment* τ is a function from the set of atomic formulae to {true, false}. We define the truth relation $\models_\tau$ relative to an assignment τ by the following induction on the structure of a formula in $\mathcal{P}^N$:

(T0): for any atomic formula A, $\models_\tau A$ iff $\tau(A) =$ true;

(T1): $\models_\tau \neg A$ iff not $\models_\tau A$;

(T2): $\models_\tau A \supset B$ iff not $\models_\tau A$ or $\models_\tau B$;

(T3): $\models_\tau \bigwedge \Phi$ iff $\models_\tau A$ for all $A \in \Phi$;

(T4): $\models_\tau \bigvee \Phi$ iff $\models_\tau A$ for some $A \in \Phi$.

The following is the standard soundness-completeness theorem: for any $A \in \mathcal{P}^N$,

$$\vdash_\mathbf{0} A \text{ if and only if } \models_\tau A \text{ for any assignment } \tau. \tag{4}$$

The only-if part is equivalent to that if there is an assignment τ such that $\models_\tau A$, then A is consistent with respect to $\vdash_\mathbf{0}$. We will refer to this as *Soundness for* $\vdash_\mathbf{0}$.[9]

[9] GL_0 is also the classical logic with the language $\mathcal{P}$ (allowing B_1 and B_2). In this case, an assignment τ is defined over the set of atomic formulae and $\mathrm{B}_i(A)$ for any A and $i = 1, 2$. That is, any formula $\mathrm{B}_i(A)$ is treated in the same as atomic formulae. Then we have the completeness-soundness result for this logic. Hence we can use any tautologies in this sense.

When Γ and A are nonepistemic, Lemma 3.2.(1) can be slightly modified,

$$\text{if } \Gamma \vdash_{\mathbf{0}} A, \text{ then } \mathrm{B}_i(\Gamma) \vdash \mathrm{B}_i(A), \tag{5}$$

which will be used without referring.

We need to prepare the *belief-elimination operator* ε : for any formula $A \in \mathcal{P}$, εA is the formula obtained from A by eliminating all the occurrences of B_1 and B_2 in A. Then the following hold (cf., Kaneko-Nagashima [7], p.340): for any subset Γ of $\mathcal{P}$ and any formula $A \in \mathcal{P}$,

$$\text{if } \Gamma \vdash A, \text{ then } \varepsilon\Gamma \vdash_{\mathbf{0}} \varepsilon A, \tag{6}$$

where $\varepsilon\Gamma = \{\varepsilon B : B \in \Gamma\}$. This will be used later.

3.2 Game Theoretical Concepts in the Formalized Language $\mathcal{P}$

Now we redescribe the game theoretical concepts given in Subsection 2.1 in $\mathcal{P}$.

We describe the payoff functions g_1, g_2 in terms of symbols R_1, R_2 as follows: for $i = 1, 2$,

(g_i): $\{R_i(a;b) : g_i(a) \geq g_i(b)\} \cup \{\neg R_i(a';b') : g_i(a') < g_i(b')\}$.

We denote this set of formulae by the same symbol g_i as the payoff function g_i. This should cause no confusions. This describes the payoff function g_i as preferences R_i.

We define the best strategy property conditional upon some strategy a_j by

$$\bigwedge_{y_i \in \Sigma_i} R_i(a_i, a_j : y_i, a_j), \tag{7}$$

which is denoted by $\mathrm{Nash}_i(a_i \mid a_j)$. We denote the following formulae by $\mathrm{Nash}_i(a_i)$ and $\mathrm{Nash}(a_1, a_2)$:

$$\bigwedge_{y_j \in \Sigma_j} \mathrm{Nash}_i(a_i \mid y_j); \text{ and } \mathrm{Nash}_1(a_1 \mid a_2) \bigwedge \mathrm{Nash}_2(a_2 \mid a_1). \tag{8}$$

That is, $\mathrm{Nash}_i(a_i)$ means that a_i is a dominant strategy for player i, and $\mathrm{Nash}(a_1, a_2)$ means that (a_1, a_2) is a Nash equilibrium. In the following, we abbreviate $\bigwedge_{y_j \in \Sigma_j}$ as $\bigwedge_{y_j}$, etc, which will cause no confusions. Since either $g_i \vdash_{\mathbf{0}} R_i(a : b)$ or $g_i \vdash_{\mathbf{0}} \neg R_i(a : b)$ for any $a, b \in \Sigma$, $\mathrm{Nash}_i(a_i \mid a_j)$ and

$\mathrm{Nash}_i(a_i)$ (or $\mathrm{Nash}(a_1, a_2)$) are decidable when g_i (or (g_1, g_2)) is assumed, e.g.,

$$\begin{array}{c} \text{if } a_i \text{ is a dominant strategy, then } g_i \vdash_{\mathbf{0}} \mathrm{Nash}_i(a_i), \\ \text{if not, then } g_i \vdash_{\mathbf{0}} \neg\mathrm{Nash}_i(a_i), \end{array} \tag{9}$$

which will be used later.

After Section 5, we assume that each player i knows (truly believes his own payoff function g_i), which is described as $\mathrm{B}_i^+(g_i)$ as a nonlogical axiom. In general, he does not know the other player j's payoff function g_j but only has a belief on it. For example, player i believes that j's payoff function is given as $\hat{g}_j$. Here we postulate that when each player i has a belief $\hat{g}_j$ on the other player's payoff function, player i assumes that j knows $\hat{g}_j$. In this case, we assume

$$\mathrm{B}_i^+(g_i), \mathrm{B}_i\mathrm{B}_j^+(\hat{g}_j)$$

Of course, we need to add at least $\mathrm{B}_j^+(g_j)$ as an nonlogical axiom. If $\hat{g}_j$ differs from g_j, this addition would lead a contradiction in epistemic logic $\mathrm{S4}^2$. However, this yields any logical problem in $\mathrm{KD4}^2$.

The following hold.

Lemma 3.3. For any strategy profile $a, b \in \Sigma$,

(1): $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(R_i(a:b)) \supset R_i(a:b)$;

(2): $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(\neg R_i(a:b)) \supset \neg R_i(a:b)$;

(3): $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid a_j)) \supset \mathrm{Nash}_i(a_i \mid a_j)$;

(4): $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(\mathrm{Nash}_i(a_i)) \supset \mathrm{Nash}_i(a_i)$.

Proof. We prove (2), and (1) can be proved similarly. First, since either $g_i \vdash R_i(a:b)$ or $g_i \vdash \neg R_i(a:b)$, we have $\mathrm{B}_i^+(g_i) \vdash R_i(a:b)$ or $\mathrm{B}_i^+(g_i) \vdash \neg R_i(a:b)$. If $\mathrm{B}_i^+(g_i) \vdash \neg R_i(a:b)$, then $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(\neg R_i(a:b)) \supset \neg R_i(a:b)$. Suppose $\mathrm{B}_i^+(g_i) \vdash R_i(a:b)$. Then $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(R_i(a:b))$. Since $\vdash \neg(\mathrm{B}_i(R_i(a:b)) \bigwedge \mathrm{B}_i(\neg R_i(a:b)))$ by Axiom D_i, we have $\vdash \mathrm{B}_i(R_i(a:b)) \supset \neg\mathrm{B}_i(\neg R_i(a:b))$. Hence $\mathrm{B}_i^+(g_i) \vdash \neg\mathrm{B}_i(\neg R_i(a:b))$. Hence $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(\neg R_i(a:b)) \supset \neg R_i(a:b)$.

We can prove (3) and (4) using (1) and Lemma 2.1.(4). ■

3.3 $\mathcal{J}$-Belief Sets for Player i

We look at an interaction structure $\mathcal{J} = (J_1, J_2)$ from the viewpoint of player i. As stated, this interaction structure describes what interpersonal

belief/knowledge is required. When player i believes that beliefs are distributed between the players such as $g^i = (g_i, \hat{g}_j)$, the $\mathcal{J}$-beliefs of player i describe what interpersonal belief/knowledge is required by $\mathcal{J} = (J_1, J_2)$.

Let $\mathcal{A} = (A_1, A_2)$ be a given vector of formulae, which we look at from player i's viewpoint. First, we define a j-formula from $\mathcal{J} = (J_1, J_2)$ and $\mathcal{A} = (A_1, A_2)$ by

(B-0): $\mathrm{B}_j^+(A_j)$ is a j-formula for any $j = 1, 2$;

(B-1): if C is a j-formula and if $j \in J_k$ and $k \neq j$, then $\mathrm{B}_k(C)$ is a k-formula.

Then we denote the set of all i-formulae by $\mathsf{B}^i(\mathcal{J}, \mathcal{A})$ generated by (B-0) and (B-1) from $\mathcal{J}$ and $\mathcal{A}$. Recall that we look at $\mathcal{J} = (J_1, J_2)$ and $\mathcal{A} = (A_1, A_2)$ from i's view point. Hence A_i is known to player i and player i believes that A_j is known to player j. Hence the i-formulae generated by (B-0) and (B-1) are what we want, and the j-formulae appearing in (B-0) and (B-1) are auxiliary concepts appearing in intermediate steps.

The following are examples of $\mathsf{B}^i(\mathcal{J}, \mathcal{A})$:

(1): For $\mathcal{J}^1 = (\{1\}, \{2\})$ and $g = (g_1, g_2)$,

$$\mathsf{B}^i(\mathcal{J}^1, g) = \{\mathrm{B}_i^+(g_i)\} \text{ for } i = 1, 2.$$

In this case, player i truly believes the payoff function g_i and does not think about the other's payoff function. Here only (B-0) is applied to $g = (g_1, g_2)$ once.

(2): For $\mathcal{J}^2 = (\{1\}, \{1, 2\})$, $g^1 = (g_1, \hat{g}_2)$ and $g^2 = (\hat{g}_1, g_2)$,

$$\mathsf{B}^1(\mathcal{J}^2, g^1) = \{\mathrm{B}_1^+(g_1)\}; \text{ and } \mathsf{B}^2(\mathcal{J}^2, g^2) = \{\mathrm{B}_2\mathrm{B}_1^+(\hat{g}_1), \mathrm{B}_2^+(g_2)\}.$$

Hence $\mathsf{B}^1(\mathcal{J}^2, g^2)$ is the same as $\mathsf{B}^1(\mathcal{J}^1, g)$ for player 1, since he ignores 2 in either interaction structures. The set $\mathsf{B}^2(\mathcal{J}^2, g^2)$ means that player 2 knows his own payoff function g_2 and believes that 1's payoff function is $\hat{g}_1$ and is known to 1. For $\mathsf{B}^2(\mathcal{J}^2, g^2)$, we apply (B-0) to each player, and then (B-1) to player 2 with $\mathrm{B}_1^+(\hat{g}_1)$.

In either case, the above inductive definition stops in one or two steps. The last case does not stop and generates an infinite set.

(4): For $\mathcal{J}^4 = (\{1, 2\}, \{1, 2\})$ and $g^1 = (g_1, \hat{g}_2)$ and $g^2 = (\hat{g}_1, g_2)$,

$$\mathsf{B}^1(\mathcal{J}^4, g^1) = \{\mathrm{B}_1^+(g_1), (\mathrm{B}_1\mathrm{B}_2)^1\mathrm{B}_1^+(g_1), (\mathrm{B}_1\mathrm{B}_2)^2\mathrm{B}_1^+(g_1), ...\} \cup$$

$$\{\mathrm{B}_1\mathrm{B}_2^+(\hat{g}_2), (\mathrm{B}_1\mathrm{B}_2)^1\mathrm{B}_1\mathrm{B}_2^+(\hat{g}_2), (\mathrm{B}_1\mathrm{B}_2)^2\mathrm{B}_1\mathrm{B}_2^+(\hat{g}_2), ...\};$$

and $\mathsf{B}^2(\mathcal{J}^4,g^2)$ is symmetrically described. The set $\mathsf{B}^1(\mathcal{J}^4,g^1)$ means that player 1 knows his payoff function g_1, and believes that 2 knows his payoff function $\hat{g}_2$, 1 believes 2 believes 1 knows g_1, and so on. In other words, 1 *believes* that $(g_1,\hat{g}_2)$ is common knowledge between 1 and 2, and his belief on g_1 is true. The second set $\mathsf{B}^2(\mathcal{J}^4,g^2)$ is symmetric to $\mathsf{B}^1(\mathcal{J}^4,g^1)$, i.e., it means that player 2 believes that $(\hat{g}_1,g_2)$ is common knowledge. In these cases, the above inductive definition does not terminates in a finite number of steps, and generates these infinite sets.

In our logic KD4^2, we can assume $\mathsf{B}^1(\mathcal{J}^4,g^1)$ and $\mathsf{B}^2(\mathcal{J}^4,g^2)$ without having a contradiction. In S4^2, the set $\mathsf{B}^1(\mathcal{J}^4,g^1)$ describes that $g_1 \bigwedge \hat{g}_2$ is common knowledge. In this case, we *cannot* allow g^1 and g^2 to be different since these two sets are inconsistent if they are different.

The inductive definition starts with $\mathrm{B}_i^+(A_i)$, but it is still capable to talk about beliefs for player i instead of knowledge. For example, if we take $\mathcal{A} = (\mathrm{B}_1(A_1),\mathrm{B}_2(A_2))$, then $\mathsf{B}^1(\mathcal{J}^1,\mathcal{A}) = \{\mathrm{B}_1^+\mathrm{B}_1(A_1)\}$, which is equivalent to $\{\mathrm{B}_1(A_1)\}$. In this sense, the above definition is flexible enough to treat beliefs and knowledge.

4 Final Decision Axioms

Let $\mathcal{J} = (J_1,J_2)$ be an interaction structure. We consider the system of symbols, $(\{I_{1j}\}_{j\in J_1},\{I_{2j}\}_{j\in J_2})$, where each I_{ij} is the unary predicate symbol introduced in Section 2. With $I_{ij}(a_j)$, as stated, we associate the meaning that player i predicts that a_j is a final decision reached by player j. Particularly, the intended meaning of $I_{ii}(a_i)$ is that a_i is a final decision made by player i himself. These intended meanings of the symbols are described by the following (nonlogical) axioms on $(\{I_{1j}\}_{j\in J_1},\{I_{2j}\}_{j\in J_2})$.

Base Axioms: for $i = 1,2$,

(I1$_i$): (Best Response Property): $\bigwedge\limits_{x}\left(\bigwedge\limits_{j\in J_i} I_{ij}(x_j) \supset \bigwedge\limits_{y_i} \mathrm{B}_i(R_i(x:y_i;x_j))\right)$;

(I2$_i$): (Belief of Predictions): $\bigwedge\limits_{j\in J_i}\bigwedge\limits_{k\in J_j}\bigwedge\limits_{x_k} (I_{ik}(x_k) \supset \mathrm{B}_i(I_{jk}(x_k)))$;

(I3$_i$): (Necessity of Predictions): $\bigwedge\limits_{j\in J_i}\bigwedge\limits_{k\in J_j}\left(\bigvee\limits_{x_j} I_{ij}(x_j) \supset \bigvee\limits_{x_k} I_{ik}(x_k)\right)$.

We denote the conjunction, $\mathrm{I1}_i \bigwedge \mathrm{I2}_i \bigwedge \mathrm{I3}_i$, of these axioms by I_i(1–3) for each $i = 1, 2$, and (I_1(1–3),I_2(1–3)) by I(1–3). Also, we denote the conjunction of $\mathrm{I1}_i$ and $\mathrm{I2}_i$ by I_i(1,2), etc.

We will consider the $\mathcal{J}$-belief set $\mathsf{B}^i(\mathcal{J}$,I(1–3)), but before it, let us consider the contents of the above base axioms:

($\mathrm{I1}_i$): if i predicts that players i and j could choose a_i and a_j, then his own choice a_i should be a best response to a_j;

($\mathrm{I2}_i$): if player k is in the scope of player j who is in the scope of player i and if i predicts that a_k is k's decision, then i believes that j also predicts the same;

($\mathrm{I3}_i$): if player i has a prediction on j's decision, then i has predictions on choices of the players in J_j.

When $J_i = \{i\}$, $\mathrm{I3}_i$ would be vacuous, which case will be discussed in Section 5

As far as J_i includes the other player j, Axiom $\mathrm{I2}_i$ differs from the other axioms in that it connects i's prediction to the other player j's prediction. It is the axiom which requires interpersonal considerations. When $J_i = \{i\}$, no interpersonal consideration is required. When $J_i = \{i, j\}$, this axiom requires player i to think about $I_{jj}(a_j)$, *a fortiori,* about I_j(1–3). For this thinking, we apply the $\mathcal{J}$-belief set $\mathsf{B}^i(\mathcal{J}$,I(1–3)) introduced in Subsection 3.3 to these axioms.

We emphasize the subjectivity of the $\mathcal{J}$-belief set $\mathsf{B}^i(\mathcal{J}$,I(1–3)). For example, when $J_i = \{i, j\}$ and $J_j = \{j\}$, we have

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)) = \{\mathrm{B}_i^+(\mathrm{I}_i(1\text{–}3)), \mathrm{B}_i\mathrm{B}_j^+(\mathrm{I}_j(1\text{–}3))\}.$$

That is, player i assumes the above axioms and believes that player j assumes the above axiom based on interaction structure $\mathcal{J} = (J_1, J_2)$.[10] In this case, Axiom $\mathrm{I2}_i$ includes $I_{jj}(a_j)$, and to determine this, player i's belief on $\mathrm{B}_j^+(\mathrm{I}_j(1\text{–}3))$ is required. Since $\mathsf{B}^i(\mathcal{J}$,I(1–3)) represents i's subjective thinking, it may be the case that player j is thinking actually based on a different interaction structure $\mathcal{J}'$, e.g., he is thinking based on $\mathsf{B}^j(\mathcal{J}'$,I(1–3))

[10] It is almost meaningless to talk about whether the behavioral axioms I_i(1–3) is knowledge or belief, following our distinction between "knowledge" and "belief". The problem here is whether or not player i adopts the axioms I_i(1–3). Unless there is a diagreement between his adoption and his belief, the axioms I_i(1–3) are true in the trivial sense. On the other hand, we can talk about the distinction between player i's knowledge and belief on the other player's axioms I_j(1–3).

(note that I(1–3) depends upon $\mathcal{J}'$) or even his thinking is totally different from these axioms.

We adopt the $\mathcal{J}$-belief set, $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3))$, for player i as a nonlogical axiom, and then would like to "solve" or "determine" $I_{ij}(\cdot)$. Recall the practice learned at *middle school for solving a simultaneous equation*:

(1^0): assuming that the equation has solutions, we derive solutions from the equation;

(2^0): we verify that the derived solution satisfy the equation.

In our case, we will take similar steps:

(1): the first is to derive some formulae as candidates for $I_{ij}(\cdot)$ from $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3))$;

(2): the second is to verify that the candidates for $I_{ij}(\cdot)$, indeed, satisfies axiom $\mathrm{I}_i(1\text{–}3)$.

We need to prepare some axiom (schema) to formulate the second step in our language.

Consider an interaction structure $\mathcal{J} = (J_1, J_2)$ and a player i. Then let

$$\mathcal{A} = (\{A_{1j}(a_j) : a_j \in \Sigma_j, j \in J_1\}, \{A_{2j}(a_j) : a_j \in \Sigma_j, j \in J_2\})$$

and we denote, by $\mathrm{wd}[\mathcal{A}]$ the vector

$$\left(\bigwedge_{j\in J_1} \bigwedge_{x_j} (A_{1j}(x_j) \supset I_{1j}(x_j)), \quad \bigwedge_{j\in J_2} \bigwedge_{x_j} (A_{2j}(x_j) \supset I_{2j}(x_j)) \right).$$

Then we consider the following axiom schemata:

$(\mathbf{WD}_i)$: $(\bigwedge \mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)[\mathcal{A}])) \supset \bigwedge \mathsf{B}^i(\mathcal{J},\mathrm{wd}[\mathcal{A}])$,

where $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)[\mathcal{A}])$ is obtained from $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3))$ by substituting $A_{lk}(a_k)$'s for all occurrences of $I_{lk}(a_k)$'s in $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3))$. The premise states that the candidates in $\mathcal{A}$ satisfy $\mathrm{I}_i(1\text{–}3)$, and WD_i states that if this is the case, then each $A_{lk}(a_k)$ implies $I_{lk}(a_k)$. This together with step (1) gives the solution to each $I_{lk}(a_k)$.

Here it is important to emphasize that when $\mathcal{J} = \mathcal{J}^4 = (\{1,2\},\{1,2\})$, we *cannot* take the conjunction of $\mathsf{B}^i(\mathcal{J},\mathrm{WD})$, since it is an infinite set. Hence the above WD_i is not available in the present $\mathrm{KD4}^2$. This is one reason that we need an infinitary extension of our $\mathrm{KD4}^2$, which will be a subject of Part II.

Again, we take the $\mathcal{J}$-belief sets of these schemata as far as WD = (WD_1,WD_2) is defined. For example, when $J_i = \{i, j\}$ and $J_j = \{j\}$, we have

$$\mathsf{B}^i(\mathcal{J}, \mathrm{WD}) = \mathrm{B}_i^+(\mathrm{WD}_i) \cup \mathrm{B}_i\mathrm{B}_j^+(\mathrm{WD}_j).$$

Recall that each WD_i is a schema and is regarded as a set here.

Now we are in a state to start solving these axioms in each particular case.

5 Interaction Structure $\mathcal{J}$ with $J_i = \{i\}$: One-Person Decision Making

Consider an interaction structure $\mathcal{J}$ with $J_i = \{i\}$ from the viewpoint of player i. As stated already, no interpersonal feature is involved and the problem is regarded as one-person decision making. Thus the problem is not game theoretic in the standard sense. However, it is the point here that player i regards his decision making as a one-person problem but his subjective thinking differs from reality.

5.1 Characterization Theorem

When player i adopts an interaction structure $\mathcal{J}$ with $J_i = \{i\}$, Axiom $\mathrm{I3}_i$ is vacuous. Axioms $\mathrm{I1}_i$ and $\mathrm{I2}_i$ are written as

$$\mathrm{I1}_i : \bigwedge_{x_i} (I_{ii}(x_i) \supset \mathrm{B}_i(\mathrm{Nash}_i(x_i))) ;$$

$$\mathrm{I2}_i : \bigwedge_{x_i} (I_{ii}(x_i) \supset \mathrm{B}_i(I_{ii}(x_i))) .$$

Thus $\mathrm{I1}_i$ means that if x_i is i's decision, then it is a dominant strategy, and $\mathrm{I2}_i$ that player i believes (*a fortiori*, knows) his decision.

In this case, the $\mathcal{J}$-belief set $\mathsf{B}^i(\mathcal{J}$,I(1–3)) is given as $\{\mathrm{B}_i^+(\mathrm{I1}_i \bigwedge \mathrm{I2}_i)\}$. The axiom schema WD_i is given as the set of formulae

$$\mathrm{B}_i^+(\mathrm{I1}_i \bigwedge \mathrm{I2}_i)[\mathcal{A}^i] \supset \bigwedge_{x_i} \mathrm{B}_i^+(A_{ii}(x_i) \supset I_{ii}(x_i)), \tag{10}$$

where $\mathcal{A}^i = \{A_{ii}(x_i)\}_{x_i \in \Sigma_i}$ is a family of formulae indexed by x_i. Then $\mathsf{B}^i(\mathcal{J}$,WD) is given as $\{\mathrm{B}_i^+ \left(\mathrm{B}_i^+(\mathrm{I1}_i \bigwedge \mathrm{I2}_i)[\mathcal{A}^i] \supset \bigwedge_{x_i} \mathrm{B}_i^+(A_{ii}(x_i) \supset I_{ii}(x_i)) \right)\}$.

This and $\mathsf{B}^i(\mathcal{J}$,I(1–3)) determine a dominant strategy behavior, which will be proved in the end of this subsection.

Theorem 5.A (Characterization I):

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J}, \mathrm{WD}) \vdash \bigwedge_{x_i} \mathrm{B}_i^+ \left(I_{ii}(x_i) \equiv \mathrm{B}_i(\mathrm{Nash}_i(x_i))\right). \tag{11}$$

Theorem 5.A is purely solution-theoretic in that it does not depend upon the belief/knowledge on the payoff function g_i. That is, as far as player i follows the final decision axioms with respect to interaction structure $\mathcal{J}$ with $J_i = \{i\}$ as his behavior principle, his decision making is to choose a dominant strategy, even though he may fail to find such a decision. It is important to notice that the conclusion of (11) has the outer B_i^+, which means that he knows this conclusion itself. To look for a specific dominant strategy, he should use his belief $\tilde{g}_i$ on his payoff function g_i. The problem of whether he could find it or not will be considered in the next subsection.

When player i knows his own payoff function g_i, then (11) can be written as

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J}, \mathrm{WD}), \mathrm{B}_i^+(g_i) \vdash \bigwedge_{x_i} \mathrm{B}_i^+ \left(I_{ii}(x_i) \equiv \mathrm{B}_i^+ \left(\mathrm{Nash}_i(x_i)\right)\right), \tag{12}$$

which will be proved in the end of this subsection. This means that if player i knows his payoff function g_i, the determined choice should be a dominant strategy in the objective sense. However, if he has a *false belief* $\tilde{g}_i$ on g_i, then $\mathrm{B}_i(\mathrm{Nash}_i(x_i))$ means i's belief of a dominant strategy *relative to his believed payoff function* $\tilde{g}_i$. When player i's belief is false, the content of his belief is incompatible with the true payoff function g_i. Then it is here a problem if the premise, $\mathsf{B}^i(\mathcal{J}$,I(1–4)), $\mathsf{B}^i(\mathcal{J}$,WD),$\mathrm{B}_i(\tilde{g}_i), g_i$, is consistent, i.e., it is not the case that

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J}, \mathrm{WD}), \mathrm{B}_i(\tilde{g}_i), g_i \vdash \ \bot. \tag{13}$$

Here $\bot$ is any contradictory formula, i.e., $\neg B \bigwedge B$. In Subsection 5.3, we will prove a deeper version of this consistency. Hence we omit the proof of the consistency of the premises of (13).

Let us prove Theorem 5.A. We denote $\mathrm{B}_i(\mathrm{Nash}_i(a_i))$ by $\hat{I}_{ii}(a_i)$. Then $\mathrm{I1}_i$ is written as

$$\mathrm{I1}_i \vdash \ I_{ii}(a_i) \supset \hat{I}_{ii}(a_i) \text{ for any } a_i \in \Sigma_i. \tag{14}$$

Here we prove the converse.

Proof of Theorem 5.A. First, we prove that the premise of (10) holds when we substitute $\{\hat{I}_{ii}(x_i)\}_{x_i}$ for $\{A_{ii}(x_i)\}_{x_i}$. The premise is written as the conjunction of the following:

$$\mathrm{B}_i^+(\mathrm{I1}_i[\{\hat{I}_{ii}(x_i)\}_{x_i}]): \quad \mathrm{B}_i^+\left(\bigwedge_{x_i}(\hat{I}_{ii}(x_i) \supset \mathrm{B}_i(\mathrm{Nash}_i(x_i)))\right);$$

$$\mathrm{B}^+(\mathrm{I2}_i[\{\hat{I}_{ii}(x_i)\}_{x_i}]): \quad \mathrm{B}_i^+\left(\bigwedge_{x_i}(\hat{I}_{ii}(x_i) \supset \mathrm{B}_i(\hat{I}_{ii}(x_i)))\right).$$

These are provable in KD4[2] without any additional assumption. Hence

$$\mathrm{B}_i^+(\mathrm{I1}_i \bigwedge \mathrm{I2}_i), \mathrm{WD}_i \vdash \ \hat{I}_{ii}(a_i) \supset I_{ii}(a_i) \text{ for all } a_i \in \Sigma_i.$$

From this and (14), we obtain $\mathsf{B}^i(\mathcal{J},\mathrm{I(1\text{–}3)}), \mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash \mathrm{B}_i^+(I_{ii}(a_i) \equiv \mathrm{B}_i(\mathrm{Nash}_i(a_i)))$ for all $a_i \in \Sigma_i$. Hence we have (11). ■

Proof of (12): Let a_i be an arbitrary strategy. It follows from Lemma 2.1 that $\mathrm{B}_i^+(g_i) \vdash \mathrm{B}_i(\mathrm{Nash}_i(a_i)) \supset \mathrm{Nash}_i(a_i)$. Since $\mathsf{B}^i(\mathcal{J},\mathrm{I(1\text{–}3)}), \mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash I_{ii}(a_i) \equiv \mathrm{B}_i(\mathrm{Nash}_i(a_i))$ by Theorem 5.A, we have $\mathsf{B}^i(\mathcal{J},\mathrm{I(1\text{–}4)}), \mathsf{B}^i(\mathcal{J},\mathrm{WD}), \mathrm{B}_i^+(g_i) \vdash I_{ii}(a_i) \equiv \mathrm{B}_i^+(\mathrm{Nash}_i(a_i))$. This implies (12) by Lemma 32.(3). ■

5.2 Decidability and Playability

We assume that player i follows the final decision axioms with respect to $\mathcal{J}$ with $J_i = \{i\}$. Then his behavior is determined by Theorem 5.A. Here we consider the problem of whether he can find, using his belief $\tilde{g}_i$ on g_i, a specific dominant strategy or not. Thus we have the following assumptions which is obtained by adding $\mathrm{B}_i(\tilde{g}_i)$ to the assumption of (11):

$$\mathsf{B}^i(\mathcal{J},\mathrm{I(1\text{–}3)}), \mathsf{B}^i(\mathcal{J},\mathrm{WD}), \mathrm{B}_i(\tilde{g}_i). \tag{15}$$

We denote the union of (15) by $\tilde{\Delta}_i$. The following theorem states that player i can decide whether or not a given strategy is a dominant strategy relative to $\tilde{g}_i$. Here we assume that $\tilde{\Delta}_i$ is consistent, which, in fact, can be proved in the same manner as in the proof of Theorem 5.D.

Theorem 5.B (Decidability I):

(0-1): a_i is a dominant strategy for $\tilde{g}_i$ if and only if $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(I_{ii}(a_i))$;

(0-2): a_i is not a dominant strategy for $\tilde{g}_i$ if and only if $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\neg I_{ii}(a_i))$.

Proof. We prove (0-2). Suppose that a_i is not a dominant strategy for $\tilde{g}_i$. The formal counterpart of this supposition is written as $\tilde{g}_i \vdash_{\mathbf{0}} \neg\mathrm{Nash}_i(a_i)$

by (9). This implies $\mathrm{B}_i(\tilde{g}_i) \vdash \mathrm{B}_i(\neg\mathrm{Nash}_i(a_i))$. From this, we have $\mathrm{B}_i(\tilde{g}_i) \vdash \neg\mathrm{B}_i(\mathrm{Nash}_i(a_i))$ by Lemma 3.1.(6). Hence $\tilde{\Delta}_i \vdash \neg I_{ii}(a_i)$. Then $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\neg I_{ii}(a_i))$ by Lemma 3.2.(3).

Suppose $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\neg I_{ii}(a_i))$. Then $\tilde{\Delta}_i \vdash \neg I_{ii}(a_i)$, i.e., $\tilde{\Delta}_i \vdash \neg\mathrm{B}_i(\mathrm{Nash}_i(a_i))$. Recall $\tilde{g}_i \vdash_0 \mathrm{Nash}_i(a_i)$ or $\tilde{g}_i \vdash_0 \neg\mathrm{Nash}_i(a_i)$. Suppose $\tilde{g}_i \vdash_0 \mathrm{Nash}_i(a_i)$. Then $\mathrm{B}_i(\tilde{g}_i) \vdash \mathrm{B}_i(\mathrm{Nash}_i(a_i))$, which means that $\tilde{\Delta}_i$ is inconsistent, a contradiction. Hence $\tilde{g}_i \vdash_0 \neg\mathrm{Nash}_i(a_i)$. This is a formal counterpart of the left-hand side of (0-2). ■

Theorem 5.B has two important implications. First, player i can find his final decisions – dominant strategies – when his believed payoff function $\tilde{g}_i$ allows a dominant strategy. This is needed to play a game as far as player i adopts the present interaction structure $\mathcal{J}$ with $J_i = \{i\}$. Second, this theorem implies that player i can decide whether or not he has a final decision, which is summarized as follows.

Theorem 5.C (Playability I):

(0-3): $\tilde{g}_i$ allows a dominant strategy if and only if $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\bigvee_{x_i} I_{ii}(x_i))$;

(0-4): $\tilde{g}_i$ allows no dominant strategy if and only if $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\bigwedge_{x_i} \neg I_{ii}(x_i))$.

The first states that when $\tilde{g}_i$ allows a dominant strategy, player i notices this fact, and Theorem 5.B means that he can decide whether each strategy is dominant. The second leads player i to a further step. When $\tilde{g}_i$ allows no dominant strategies, (0-2) states that player i infers that each strategy is not his final decision, and (0-4) implies that he recognizes that he has no final decision. In this case, *he may notice the need of thinking about the other player's decision making*, and then may go to another interaction structure. This will be discussed in Section 6 and Part II.

Proof of Theorem 5.C. We prove only (0-3). Suppose that a_i is a dominant strategy for $\tilde{g}_i$. This is stated as $\tilde{g}_i \vdash_0 \mathrm{Nash}_i(a_i)$. Then $\mathrm{B}_i(\tilde{g}_i) \vdash \mathrm{B}_i(\mathrm{Nash}_i(a_i))$. Hence $\mathrm{B}_i(\tilde{g}_i) \vdash \bigvee_{x_i} \mathrm{B}_i(\mathrm{Nash}_i(x_i))$. Then it follows from Theorem 5.A that $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}4)), \mathsf{B}^i(\mathcal{J},\mathrm{WD}), \mathrm{B}_i(\tilde{g}_i) \vdash \bigvee_{x_i} I_{ii}(x_i)$. Thus $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\bigvee_{x_i} I_{ii}(x_i))$.

Conversely, suppose $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(\bigvee_{x_i} I_{ii}(x_i))$. Then $\tilde{\Delta}_i \vdash \bigvee_{x_i} I_{ii}(x_i)$. Hence by Theorem 5.A, we have $\tilde{\Delta}_i \vdash \bigvee_{x_i} \mathrm{B}_i(\mathrm{Nash}_i(x_i))$. Recall $\tilde{g}_i \vdash_0 \mathrm{Nash}_i(a_i)$ or

$\tilde{g}_i \vdash_{\mathbf{0}} \neg\mathrm{Nash}_i(a_i)$ for each $a_i \in \Sigma_i$. Suppose $\tilde{g}_i \vdash_{\mathbf{0}} \neg\mathrm{Nash}_i(a_i)$ for all $a_i \in \Sigma_i$. Then $\mathrm{B}_i(\tilde{g}_i) \vdash \neg\mathrm{B}_i(\mathrm{Nash}_i(a_i))$ for all a_i, i.e., $\mathrm{B}_i(\tilde{g}_i) \vdash \neg \bigvee_{x_i} \mathrm{B}_i(\mathrm{Nash}_i(x_i))$. Hence $\tilde{\Delta}_i$ is inconsistent, which is not the case. Hence $\tilde{g}_i \vdash_{\mathbf{0}} \mathrm{Nash}_i(a_i)$ for some $a_i \in \Sigma_i$. This is a formal counterpart of the left-hand side of (1). ■

5.3 Mutually False Beliefs: From the Inductive Point of View

Here we will examine the results of the above subsections from the viewpoint of inductive game theory. Recall that in an inductively stable stationary state $a^* = (a_1^*, a_2^*)$, each player i has the active and passive experiences $\mathcal{E}(i \mid a^*) = \{[(a_i; a_j), g_i(a_i; a_j)] : a_i = a_i^* \text{ or } a_j = a_j^*\}$. Here we consider the case where he has no additional information to $\mathcal{E}(i \mid a^*)$. Then these experiences tell him about his payoff function g_i up to the experienced domain $\{(a_i; a_j) : a_i = a_i^* \text{ or } a_j = a_j^*\}$, and he fills up the payoff function over the unexperienced domain. Let $\tilde{g}_i$ be a possible payoff function on Σ. We say that $\tilde{g}_i$ is *compatible with* $\mathcal{E}(i \mid a^*)$ iff

$$\tilde{g}_i(a_i; a_j) = g_i(a_i; a_j) \text{ for all } [(a_i; a_j), g_i(a_i; a_j)] \in \mathcal{E}(i \mid a^*). \tag{16}$$

Here we consider the possibilities of making false beliefs over the unexperienced domains as well as of becoming conscious of this falsity.

In the following theorem, we assume the set of axioms in (15) and the true payoff function g_i, i.e.,

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J}, \mathrm{WD}), \mathrm{B}_i(\tilde{g}_i), g_i \tag{17}$$

whose union is denoted by $\Delta_i(\tilde{g}_i)$. Here we ask whether or not player can construct a payoff function $\tilde{g}_i$ compatible with his experiences $\mathcal{E}(i \mid a^*)$ following his final decision axiom with interaction structure $\mathcal{J}$ with $J_i = \{i\}$. The following theorem states that as far as he takes only $\mathcal{E}(i \mid a^*)$ into account, this is always possible, and moreover, that both players can follow simultaneously this thought process without yielding a logical inconsistency.

Theorem 5.D (Inductive Pitfall I): Let $a^* = (a_1^*, a_2^*)$ be an inductively stable stationary state where it is a strict Nash equilibrium, i.e.,

$$(\mathrm{SN}) : g_i(a_i^*; a_j^*) > g_i(b_i; a_j^*) \text{ for all } b_i \neq a_i^* \text{ and } i = 1, 2.$$

Then there are payoff functions $\tilde{g}_1$ and $\tilde{g}_2$ such that

(1): $\tilde{g}_i$ is compatible with $\mathcal{E}(i \mid a^*)$ for $i = 1, 2$;

(2): each $\tilde{g}_i$ allows a dominant strategy;

(3): each i has interaction structure $\mathcal{J}$ with $J_i = \{i\}$;

(4): $\Delta_i(\tilde{g}_i) \vdash I_{ii}(a_i^*)$ and $\Delta_i(\tilde{g}_i) \vdash \neg I_{ii}(a_i)$ for all $a_i \neq a_i^*$ and $i = 1, 2$;

(5): $\Delta_1(\tilde{g}_1) \cup \Delta_2(\tilde{g}_2)$ is consistent in $\mathrm{KD4}^2$.

Thus, if each player i knows his payoff function g_i up to $\mathcal{E}(i \mid a^*)$, then he may construct complete preferences by filling up the other part in the way compatible with his experiences $\mathcal{E}(i \mid a^*)$. Theorem 5.D claims that both players can follow this process simultaneously without yielding an inconsistency. It is the point of the theorem that $\tilde{g}_1$ and $\tilde{g}_2$ may be very different from true g_1 and g_2. For example, the game $g = (g_1, g_2)$ can be the game of Table 1.3, where neither player has a dominant strategy. However, it is possible for each player to construct $\tilde{g}_i$ from his experiences $\mathcal{E}(i \mid a^*)$ so that it explains his behavior by the dominant strategy behavior principle.

In epistemic logic $\mathrm{S4}^2$, the above theorem would fail without giving some restriction on $g = (g_1, g_2)$, since $\tilde{g}_i$ must be true, i.e., is g_i, in $\mathrm{S4}^2$. Hence our distinction between belief and knowledge in $\mathrm{KD4}^2$ is crucial in Theorem 5.D.

Theorems 5.B and 5.C are relevant for the consideration of Theorem 5.D. For example, (0-1) is extended into

(0-1*): a_i is a dominant strategy for $\tilde{g}_i$ if and only if $\tilde{\Delta}_i \vdash \mathrm{B}_i^+(I_{ii}(a_i))$
if and only if $\mathrm{B}_i(\tilde{\Delta}_i) \vdash \mathrm{B}_i\mathrm{B}_i^+(I_{ii}(a_i))$.

That is, he is conscious about (0-1) to (0-4). On the other hand, player i cannot know the entire assertion of Theorem 5.D in our framework, since the consistency claim cannot be formulated inside our epistemic logic $\mathrm{KD4}^2$. Hence if player i is satisfied by succeeding in finding a payoff function $\tilde{g}_i$ to explain his behavior in the sense of (4), he would get stuck to the false belief. We call this the *pitfall of induction*.

The problem arising from the consistency of $\Delta_i(\tilde{g}_i)$ cannot directly be noticed by player i, but is virtually noticeable. Player i may notice a lot of other candidates for his belief different from $\tilde{g}_i$ which are inconsistent with each other. Hence each player may notice that he has constructed a false belief. In this sense, the above pitfall would not be very serious if he is cautious enough.

Let us prove Theorem 5.D. Now we construct two believed payoff functions $\tilde{g}_1$ and $\tilde{g}_2$ satisfying the assertions of Theorem 5.D. In the case of the game of Table 1.2, $\tilde{g}_1$ may be g_1 itself, and $\tilde{g}_2$ may the payoff function

of player 2 in the game of Table 1.1 (Prisoner's Dilemma). In this case, player 2 has the false belief that his payoff function is the same as his in the Prisoner's Dilemma. This observation can be generalized in the following manner.

We can assume that $g_i(a_i;a_j) > 0$ for all $(a_1,a_2) \in \Sigma$. Let us construct $\tilde{g}_i$ as follows:

$$\tilde{g}_i(a_i,a_j) = \begin{cases} g_i(a_i,a_j) & \text{if } a_i = a_i^* \text{ or } a_j = a_j^* \\ \\ 0 & \text{otherwise.} \end{cases}$$

Then if $a_j \neq a_j^*$, then $\tilde{g}_i(a_i^*;a_j) > 0 = \tilde{g}_i(a_i;a_j)$ for all $a_i \neq a_i^*$, and $\tilde{g}_i(a_i^*;a_j^*) = g_i(a_i^*;a_j^*) \geq g_i(a_i;a_j^*) = \tilde{g}_i(a_i;a_j^*)$ for all $a_i \neq a_i^*$. Hence a_i^* is a unique dominant strategy for this $\tilde{g}_i$. Hence (1), (2) hold, and (4) follows from this construction and Theorem 5.B. Since (3) is not an assertion, we need to prove (5), now.

Here we refer to one theorem which is crucial here and later. We say that a formula A is *indecomposable* iff it is atomic or is written as $\mathrm{B}_i(C)$ for some i and C. We say that A is an B_i-formula $(i = 1,2)$ iff the outermost symbol of any maximal indecomposable subformula of A is B_i. The following theorem is proved for game logic GL_ω in Kaneko-Nagashima [8](p.281, Theorem 3.3), which can be simplified for $\mathrm{KD4}^2$.

Theorem 5.E (Separation for $\mathbf{KD4}^2$) : Let Γ_0 be a set of nonepistemic formulae, and Γ_i a set of B_i-formulae for $i = 1,2$. Let A_ℓ be a nonepistemic formula if $\ell = 0$, and A_ℓ a B_ℓ-formula if $\ell = 1$ or 2. If $\Gamma_0, \Gamma_1, \Gamma_2 \vdash A_\ell$, then $\Gamma_\ell \vdash A_\ell$ or $\Gamma_k \vdash \bot$ for some $k \neq \ell$.

Proof of Theorem 5.D. Suppose, on the contrary, that $\Delta_1(\tilde{g}_1) \cup \Delta_2(\tilde{g}_2)$ is inconsistent in $\mathrm{KD4}^2$. Then we can assume

$$\Delta_1(\tilde{g}_1) \cup \Delta_2(\tilde{g}_2) \vdash \neg R_1(a:b) \bigwedge R_1(a:b), \tag{18}$$

where a and b are any strategic combinations. Observe that the left-hand side of (18) has no occurrences of $I_{11}(\cdot)$ and $I_{22}(\cdot)$, but $I_{11}(\cdot)$ and $I_{22}(\cdot)$ occur only in $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3))$ and $\mathsf{B}^i(\mathcal{J},\mathrm{WD})$ $(i = 1,2)$. We can prove that these sets are irrelevant in (18).

Lemma 5.1. If (18) holds, then

$$\mathrm{B}_1(\tilde{g}_1), g_1, \mathrm{B}_2(\tilde{g}_2), g_2 \vdash \neg R_1(a:b) \bigwedge R_1(a:b). \tag{19}$$

Proof. Let P be a proof of (18). We can substitute $\hat{I}_{11}(a_1)$ and $\hat{I}_{22}(a_2)$ for all occurrences of $I_{11}(a_1)$ and $I_{22}(a_1)$ in P, and obtain a proof of the following:

$$\bigcup_i \{\hat{\mathsf{B}}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \hat{\mathsf{B}}^i(\mathcal{J}, \mathrm{WD}), \mathrm{B}_i(\tilde{g}_i), g_i\} \vdash \neg R_1(a:b) \bigwedge R_2(a:b). \qquad (20)$$

where $\hat{\mathsf{B}}^i(\mathcal{J},\mathrm{I}(1\text{–}3)), \hat{\mathsf{B}}^i(\mathcal{J},\mathrm{WD})$ are obtained from $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J},\mathrm{WD})$ by such substitutions. However, we can prove that $\vdash A$ for any formula A in $\hat{\mathsf{B}}^i(\mathcal{J},\mathrm{I}(1\text{–}3)) \cup \hat{\mathsf{B}}^i(\mathcal{J},\mathrm{WD})$, which is essentially the same as the two steps of the proof of Theorem 5.A. Hence (20) becomes (19). ■

Now we return to the proof of Theorem 5.D. Applying Theorem 5.E to (19), we have either

$$g_1, g_2 \vdash \neg R_1(a:b) \bigwedge R_1(a:b). \qquad (21)$$

$$\text{or } \mathrm{B}_i(\tilde{g}_i) \vdash \neg R_1(a:b) \bigwedge R_1(a:b) \text{ for at least one of } i = 1, 2. \qquad (22)$$

Applying the belief-elimination operator ε to these, we have, by (6), either $g_1, g_2 \vdash_{\mathbf{0}} \neg R_1(a:b) \bigwedge R_1(a:b)$ or $\tilde{g}_i \vdash_{\mathbf{0}} \neg R_1(a:b) \bigwedge R_1(a:b)$ for at least one of $i = 1, 2$. We prove only that the first is not the case.

We can prove that neither is the case, using Soundness for $\vdash_{\mathbf{0}}$. Consider g_1, g_2. This is simply a set of formulae which is characterized by the property that for each $i = 1, 2$ and $c, d \in \Sigma$, either $R_i(c:d) \in g_i$ or $\neg R_i(c:d) \in g_i$. We can construct an assignment σ over the atomic formulae as follows:

$$\sigma(R_i(c:d)) = \begin{cases} \text{true} & \text{if } R_i(c:d) \in g_i \\ \\ \text{false} & \text{if } \neg R_i(c:d) \in g_i \end{cases}$$

and σ is arbitrary over the other atomic formulae. By Soundness for $\vdash_{\mathbf{0}}$, the set g_1, g_2 is consistent. Hence it is not the case that $g_1, g_2 \vdash_{\mathbf{0}} \neg R_1(a:b) \bigwedge R_1(a:b)$. ■

6 Interaction Structure $\mathcal{J} = (\{i, j\}; \{j\})$: Regarding the Other Player as a One-Person Decision Maker

In this section and Part II, we replace Postulate 5 by the following:

Postulate 5^* : Each player i is cautious enough not to ignore the events with small frequencies, and hence has recorded the experiences $\{[(a_i; a_j), g_i(a_i; a_j)] : (a_i; a_j) \in \Sigma_i \times \Sigma_j\}$.

Under this postulate, it is natural to assume that his belief $\tilde{g}_i$ coincides with the true one g_i. Now, we consider the case where g_i allows no dominant strategies. According to the results of Subsection 5.2, he can notice that no dominant strategies are allowed for himself, and, moreover, may notice the need to consider player j's decision making. In this section, we assume that player i regards j as a one-person decision maker in the sense of Section 5. Thus, he uses interaction structure $\mathcal{J} = (\{i, j\}; \{j\})$ for his thought on the decision making of player j. Although player i himself finds his own payoff function, it is totally different to think about the other player's payoff function. Here player i constructs a belief $\hat{g}_j$ on player j' payoff function g_j to explain his and the other's observed behavior.

In the case of $\mathcal{J} = (\{i, j\}; \{j\})$, the arguments of Subsections 5.1 and 5.2 are applied to player j. Now, player i is thinking in the same manner as ours, and then, using the conclusions from his thinking, he makes his own decision. In Subsection 6.3, we give a result parallel to Theorem 5.D in the present case, i.e., either player has a false belief on the other's payoff function but their resulting decisions are compatible with their experiences.

Now, interpersonal features of decision making appear, and some restrictions on $g^i = (g_i; \hat{g}_j)$ become relevant: for i, j $(i \neq j)$,

(Conc$_i$): if a_j, b_j are dominant strategies for $\hat{g}_j$ and if a_i is a best response to a_j in g_i,

then a_i is a best response also to b_j for g_i.

The games of Tables 1.1–1.3 satisfy this condition. If $\hat{g}_j$ allows no dominant strategies, then Conc$_i$ holds in the trivial sense.

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$
$\mathbf{s}_{11}$	$(1,1)^*$	$(0,0)$
$\mathbf{s}_{12}$	$(1,0)$	$(0,1)^*$

Table 6.1

The game of Table 6.1 does not satisfies Conc$_i$. In this game, $\mathbf{s}_{11}$ and $\mathbf{s}_{12}$ are indifferent for player 1 and both are dominant strategies for him. However, the best strategy for 2 depends upon 1's choice. Hence 2 cannot make a decision without any additional information telling which is actually chosen by player 1. We restrict our considerations to pairs of payoff functions

satisfying this conditions in this paper.[11] We consider games without Conc_i in a separate paper.

6.1 Player i's Thinking about j's Decision Making

For interaction structure $\mathcal{J} = (\{i,j\};\{j\})$, the results of Section 5 are applied to player j. Player i's belief on Theorem 5.A is written as

$$\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{I}(1\text{–}3)),\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{WD}) \vdash \mathrm{B}_i\mathrm{B}_j^+\left(\bigwedge_{x_j}(I_{jj}(x_j) \equiv \hat{I}_{jj}(x_j))\right), \qquad (23)$$

where $\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{I}(1\text{–}3)) = \{\mathrm{B}_i\mathrm{B}_j^+(\mathrm{I}1_j \bigwedge \mathrm{I}2_j)\}$, etc. and $\hat{I}_{jj}(a_j)$ is $\mathrm{B}_j^+(\mathrm{Nash}_j(a_j))$. That is, player i believes that j behaves following the final decision axioms, $\mathrm{I}_j(1\text{–}3) \equiv \mathrm{I}1_j \bigwedge \mathrm{I}2_j$, and that his resulting behavior is to choose a dominant strategy. This follows Theorem 5.A, using Lemma 3.2.(1).

Since $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)),\mathsf{B}^i(\mathcal{J},\mathrm{WD})$ include $\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{I}(1\text{–}3)),\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{WD})$ for $\mathcal{J} = (\{i,j\};\{j\})$, it follows from (23) that

$$\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)),\mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash \mathrm{B}_i\mathrm{B}_j^+\left(\bigwedge_{x_j}(I_{jj}(x_j) \equiv \hat{I}_{jj}(x_j))\right). \qquad (24)$$

However, since (23) is more directly related to the results of Section 5 than (24), we use (23) here.

The other results of Section 5 with respect to i' belief $\hat{g}_j$ are also available for player i. We denote, by $\mathrm{B}_i(\Delta_j(\hat{g}_j))$, the union of $\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{I}(1\text{–}3))$, $\mathrm{B}_i\mathsf{B}^j(\mathcal{J},\mathrm{WD}),\mathrm{B}_i\mathrm{B}_j(\hat{g}_j)$. Then Theorems 5.B and 5.C for player i in the present case are as follows:

(1-1): a_j is a dominant strategy for $\hat{g}_j$ if and only if $\mathrm{B}_i(\Delta_j(\hat{g}_j)) \vdash \mathrm{B}_i(I_{jj}(a_j))$;

(1-2): a_j is not a dominant strategy for $\hat{g}_j$ if and only if $\mathrm{B}_i(\Delta_j(\hat{g}_j)) \vdash \mathrm{B}_i(\neg I_{jj}(a_j))$;

(1-3): $\hat{g}_j$ allows a dominant strategy if and only if $\mathrm{B}_i(\Delta_j(\hat{g}_j)) \vdash \mathrm{B}_i(\bigvee_{x_j} I_{jj}(x_j))$;

[11] This is somewhat similar to the interchangeability condition given by Nash [11], which will play a parallel role in Section 5 of Part II. In general, however, Condition Conc_i is independent of the interchangeability, which is discussed in Kaneko [3].

(1-4): $\hat{g}_j$ allows no dominant strategies if and only if $\mathrm{B}_i(\Delta_j(\hat{g}_j)) \vdash \mathrm{B}_i(\bigwedge_{x_j} \neg I_{jj}(x_j))$.

Thus, player i infers that player j can make a decision with respect to the believed payoff function $\hat{g}_j$. Note that each is further equivalent to that player i believes that player j knows the conclusion. For example, the right-hand side of (1-4) is equivalent to

$$\mathrm{B}_i(\Delta_j(\hat{g}_j)) \vdash \mathrm{B}_i\mathrm{B}_j^+(\bigwedge_{x_j} \neg I_{jj}(x_j)).$$

This states that player i notices that player j knows that he cannot make a decision. Assertions (1-1)–(1-4) are proved similarly to Theorems 5.B and 5.C.

Note that $\mathrm{B}_i(\Delta_j(\hat{g}_j))$ is included in $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J},\mathrm{WD}), \mathsf{B}^i(\mathcal{J},(g_i;\hat{g}_j))$. In fact, we can replace the former by the latter in the right-hand sides of (1-1)–(1-4).

If the left-hand side of (1-4) is the case, player i notices that player j cannot make a decision as far as he follows the dominant strategy behavior. Then player i may infer the need to go to another interaction structure, which is the subject of Part II. In the following subsection, assuming that the left-hand side of (1-3) is the case, we consider the decision making for player j with $\mathcal{J} = (\{i,j\};\{j\})$.

6.2 Player i's Own Decision Making

Consider the decision making of player i with interaction structure $\mathcal{J} = (\{i,j\};\{j\})$. In this case, player i is assumed to believe that player j knows axioms $\mathrm{I}_j(1\text{–}3) \equiv \mathrm{I1}_j \bigwedge \mathrm{I2}_j$, and his own axioms $\mathrm{I1}_i - \mathrm{I3}_i$:

$$\mathrm{I1}_i : \bigwedge_x (I_{ii}(x_i) \bigwedge I_{ij}(x_j) \supset \mathrm{Nash}_i(x_i \mid x_j)) ;$$

$$\mathrm{I2}_i : \left(\bigwedge_{k\in\{i,j\}} \bigwedge_{x_k} (I_{ik}(x_k) \supset \mathrm{B}_i(I_{ik}(x_k))) \right) \bigwedge \left(\bigwedge_{x_j} (I_{ij}(x_j) \supset \mathrm{B}_i(I_{jj}(x_j))) \right)$$

$$\mathrm{I3}_i : \bigvee_{x_i} I_{ii}(x_i) \supset \bigvee_{x_j} I_{ij}(x_j).$$

Notice that $\mathrm{I2}_i$ includes $I_{jj}(\cdot)$. To determine $I_{jj}(\cdot)$, we adopted $\mathrm{B}_j^+(\mathrm{I1}_j \bigwedge \mathrm{I2}_j)$ together with $\mathrm{B}_j^+(\mathrm{WD}_j)$. We assume that player i believes these axioms for

the determination of $I_{jj}(\cdot)$. In fact, this is taken care of by the sets $\mathsf{B}^i(\mathcal{J}$,I(1–3)) and $\mathsf{B}^i(\mathcal{J}$,WD). That is, $\mathsf{B}^i(\mathcal{J}$,I(1–3)) is given as $\{\mathrm{B}_i\mathrm{B}_j^+(\mathrm{I1}_j \bigwedge \mathrm{I2}_j), \mathrm{B}_i^+(\mathrm{I}_i(1\text{–}3))\}$. The set $\mathsf{B}^i(\mathcal{J}$,WD) is the union of $\{\mathrm{B}_i\mathrm{B}_j^+(\mathrm{WD}_j)\}$ and the set consisting of

$$\mathrm{B}_i^+\left(\bigwedge \mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)[\mathcal{A}]) \supset \bigwedge \mathsf{B}^i(\mathcal{J}, \mathrm{wd}[\mathcal{A}])\right),$$

where $\mathcal{A} = (\{A_{ik}(x_k) : x_k \in \Sigma_k \text{ and } k = i, j\}; \{A_{jj}(x_l) : x_l \in \Sigma_j\})$.

Using (24) together with the axioms for player i, we can determine $I_{ij}(a_j)$ and $I_{ii}(a_i)$ to be, respectively,

$$\mathrm{B}_i\mathrm{B}_j^+(\mathrm{Nash}_j(a_j))\left(= \mathrm{B}_i(\hat{I}_{jj}(a_j))\right); \text{ and } \bigvee_{x_j}\left(\hat{I}_{ij}(x_j) \bigwedge \mathrm{B}_i^+(\mathrm{Nash}_i(x_i \mid x_j))\right),$$

which are denoted by $\hat{I}_{ij}(a_j)$ and $\hat{I}_{ii}(a_i)$. A proof of Theorem 6.A will be given in the end of this subsection.

Theorem 6.A.(Characterization II)(1):

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J}, \mathrm{WD}) \vdash \mathrm{B}_i^+\left(\bigwedge_{x_i}(I_{ij}(x_j) \equiv \hat{I}_{ij}(x_j))\right). \tag{25}$$

(2): Let $g^i = (g_i; \hat{g}_j)$ satisfy condition Conc_i. Then

$$\mathsf{B}^i(\mathcal{J}, \mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J}, \mathrm{WD}), \mathsf{B}^i(\mathcal{J}, g^i) \vdash \mathrm{B}_i^+\left(\bigwedge_{x_i}(I_{ii}(x_i) \equiv \hat{I}_{ii}(x_i))\right). \tag{26}$$

First, we remark that when $g^i = (g_i; \hat{g}_j)$ does not satisfy Conc_i, e.g., player has the game of Table 1.3 in his mind, the second assertion does not hold, though the first holds without any condition. In a separate paper, we consider what would happen and would be required without Conc_i.

As already stated in Subsection 6.1, player i can infer whether or not player j can make a decision when j follows the dominant strategy behavior principle. We consider whether or not i can make a decision. Under our assumption, player i can make a decision if and only if $\hat{g}_j$ allows a dominant strategy. Moreover, player j can decide whether or not a given strategy is his final decision.

We denote $\mathsf{B}^i(\mathcal{J}$,I(1–3)), $\mathsf{B}^i(\mathcal{J}$,WD), $\mathsf{B}^i(\mathcal{J}, g^i)$ by $\Lambda_i(g^i)$. Note that $\Lambda_i(g^i)$ includes $\mathrm{B}_i(\Delta_j(\hat{g}_j))$ and (1-1)–(1-4) hold.

Theorem 6.B (Decidability II): Let $g^i = (g_i; \hat{g}_j)$ satisfy condition Conc_i. Then

(2-1): a_i is a best response to some dominant strategy for $\hat{g}_j$ if and only if $\Lambda_i(g^i) \vdash \mathrm{B}_i^+(I_{ii}(a_i))$;

(2-2): a_i is not a best response to any dominant strategy for $\hat{g}_j$ if and only if

$\Lambda_i(g^i) \vdash \mathrm{B}_i^+(\neg I_{ii}(a_i))$.

Theorem 6.C (Playability II): Let $g^i = (g_i; \hat{g}_j)$ satisfy condition Conc_i. Then

(2-3): $\hat{g}_j$ allows a dominant strategy if and only if $\Lambda_i(g^i) \vdash \mathrm{B}_i^+(\bigvee_{x_i} I_{ii}(x_i))$.

(2-4): $\hat{g}_j$ allows no dominant strategies if and only if $\Lambda_i(g^i) \vdash \mathrm{B}_i^+(\neg \bigvee_{x_i} I_{ii}(x_i))$.

After all, when player i believes that j's payoff function is $\hat{g}_j$ which allows a dominant strategy and which together with g_i satisfies Conc_i, player i does have a final decision which is decidable by him. Also, when $\hat{g}_j$ allows no dominant strategies, player i notices that player j could not make a decision without thinking about i's decision making, and concludes that player i himself does not have a decision. Then player i may realize the need for player j to think about i's thinking, which is the subject of Part II.

Now let us prove Theorem 6.A. The assertion (1) follows from Lemma 6.1.

Lemma 6.1. $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash I_{ij}(a_j) \equiv \hat{I}_{ij}(a_j)$.

Proof. Since $\mathrm{I2}_i \vdash I_{ij}(a_j) \supset \mathrm{B}_i(I_{jj}(a_j))$ and $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash \mathrm{B}_i(I_{jj}(a_j)) \equiv \mathrm{B}_i(\hat{I}_{jj}(a_j))$ by (24), we have $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)), \mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash I_{ij}(a_j) \supset \hat{I}_{ij}(a_j)$. We should prove the opposite.

Define $\mathcal{A} = (\{A_{ik}(x_k)\}_{x_k, k \in J_i}; \{A_{jj}(x_k)\}_{x_k})$ by

$$A_{kl}(a_l) = \begin{cases} \hat{I}_{jj}(a_j) & \text{if } k = l = j \\ \hat{I}_{ij}(a_j) & \text{if } k = i \text{ and } l = j \\ \bot & \text{otherwise,} \end{cases}$$

where $\bot$ is any contradictory formula. Then $\vdash \bigwedge(\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3))[\mathcal{A}])$, since $\vdash \mathrm{I1}_j \bigwedge \mathrm{I2}_j[\{\hat{I}_{jj}(x_j)\}_{x_j}]$ by the previous section and $\vdash (\mathrm{I1}_i \bigwedge \mathrm{I3}_i)[\mathcal{A}]$ becomes trivial because of plugging $\bot$ to the occurrences of $I_{ii}(a_i)$ and $\vdash (\mathrm{I2}_i)[\mathcal{A}]$ follows from the definition of $\hat{I}_{ij}(a_j)$. Hence $\mathsf{B}^i(\mathcal{J},\mathrm{WD}) \vdash \hat{I}_{ij}(a_j) \supset I_{ij}(a_j)$. ■

Lemma 6.2. Let Γ be a set of formulae including $\mathrm{I3}_i$. Then if $\Gamma \vdash I_{ii}(a_i) \bigwedge I_{ij}(a_j) \supset \hat{I}_{ii}(a_i)$, then $\Gamma \vdash I_{ii}(a_i) \supset \hat{I}_{ii}(a_i)$.

Proof. Suppose $\Gamma \vdash I_{ii}(a_i) \bigwedge I_{ij}(a_j) \supset \hat{I}_{ii}(a_i)$. Then $\Gamma \vdash I_{ij}(a_j) \supset \Big(I_{ii}(a_i) \supset \hat{I}_{ii}(a_i)\Big)$. Since the latter part of the assertion does not include a_j, $\Gamma \vdash \bigvee_{x_j} I_{ij}(x_j) \supset \Big(I_{ii}(a_i) \supset \hat{I}_{ii}(a_i)\Big)$. Since $\mathrm{I3}_i \vdash I_{ii}(a_i) \supset \bigvee_{x_j} I_{ij}(x_j)$, we have $\Gamma \vdash I_{ii}(a_i) \supset \hat{I}_{ii}(a_i)$. ■

Lemma 6.3. $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)) \vdash I_{ii}(a_i) \supset \hat{I}_{ii}(a_i)$.

Proof. By Lemma 6.2, it suffices to show $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)) \vdash I_{ii}(a_i) \bigwedge I_{ij}(a_j) \supset \hat{I}_{ii}(a_i)$. First, since $\mathrm{I2}_i \vdash I_{ij}(a_j) \supset \mathrm{B}_i(I_{jj}(a_j))$ and $\mathrm{I2}_j \vdash I_{jj}(a_j) \supset \mathrm{B}_j(\mathrm{Nash}_j(a_j))$, *a fortiori*, $\mathrm{B}_i(\mathrm{I2}_j) \vdash \mathrm{B}_i(I_{jj}(a_j)) \supset \mathrm{B}_i\mathrm{B}_j(\mathrm{Nash}_j(a_j))$, we have

$$\mathrm{I2}_i, \mathrm{B}_i(\mathrm{I2}_j) \vdash I_{ij}(a_j) \supset \mathrm{B}_i\mathrm{B}_j(\mathrm{Nash}_j(a_j)).$$

This together with $\mathrm{I1}_i \vdash I_{ii}(a_i) \bigwedge I_{ij}(a_j) \supset \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid a_j))$ implies

$$\mathrm{I1}_i, \mathrm{I2}_i, \mathrm{B}_i(\mathrm{I2}_j) \vdash I_{ij}(a_j) \bigwedge I_{ij}(a_j) \supset \mathrm{B}_i(\hat{I}_{jj}(a_j)) \bigwedge \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid a_j)).$$

Hence $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)) \vdash I_{ii}(a_i) \bigwedge I_{ij}(a_j) \supset \hat{I}_{ij}(a_j) \bigwedge \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid a_j))$, i.e., $\mathsf{B}^i(\mathcal{J},\mathrm{I}(1\text{–}3)) \vdash I_{ii}(a_i) \bigwedge I_{ij}(a_j) \supset \hat{I}_{ii}(a_i)$. ■

Lemma 6.4. Let $\mathcal{A} = (\{\hat{I}_{ik}(x_k)\}_{x_k, k \in J_i}; \{\hat{I}_{jj}(x_k)\}_{x_k})$. Then $\mathsf{B}^i(\mathcal{J}, g^i) \vdash \mathrm{B}_i^+(\mathrm{I}_i(1\text{–}3)[\mathcal{A}])$.

Proof. It is straightforward to verify $\vdash (\mathrm{I2}_i \bigwedge \mathrm{I3}_i)[\mathcal{A}]$.

We prove $\mathsf{B}^i(\mathcal{J}, g^i) \vdash \hat{I}_{ij}(a_j) \bigwedge \hat{I}_{ii}(a_i) \supset \mathrm{Nash}_i(a_i \mid a_j)$. Since $g^i = (g_i; \hat{g}_j)$ satisfies Conc_i, we have $g^i \vdash \mathrm{Nash}_j(a_j) \bigwedge \mathrm{Nash}_i(a_i \mid a_j) \bigwedge \mathrm{Nash}_j(b_j) \supset \mathrm{Nash}_i(a_i \mid b_j)$, which implies

$$g^i \vdash \mathrm{B}_j^+(\mathrm{Nash}_j(a_j)) \bigwedge \mathrm{Nash}_i(a_i \mid a_j) \bigwedge \mathrm{B}_j^+(\mathrm{Nash}_j(b_j)) \supset \mathrm{Nash}_i(a_i \mid b_j).$$

From this, we have $\mathrm{B}_i(g^i) \vdash \mathrm{B}_i\mathrm{B}_j^+(\mathrm{Nash}_j(a_j)) \bigwedge \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid a_j)) \bigwedge \mathrm{B}_i\mathrm{B}_j^+(\mathrm{Nash}_j(b_j)) \supset \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid b_j))$. This implies

$$\mathrm{B}_i(g^i) \vdash \Big(\hat{I}_{ij}(a_j) \bigwedge \mathrm{B}_i^+(\mathrm{Nash}_i(a_i \mid a_j))\Big) \bigwedge \hat{I}_{ij}(b_j) \supset \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid b_j))$$

Hence

$$\mathrm{B}_i(g^i) \vdash \bigvee_{x_j} \Big(\hat{I}_{ij}(x_j) \bigwedge \mathrm{B}_i^+(\mathrm{Nash}_i(a_i \mid x_j))\Big) \bigwedge \hat{I}_{ij}(b_j) \supset \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid b_j)),$$

which is $\mathrm{B}_i(g^i) \vdash \hat{I}_{ii}(a_i) \bigwedge \hat{I}_{ij}(b_j) \supset \mathrm{B}_i(\mathrm{Nash}_i(a_i \mid b_j))$. ■

Proof of Theorem 6.B. We prove only (2-1). Suppose that a_i is a best response to a dominant strategy a_j for $\hat{g}_j$. Then $\Lambda_i(g^i) \vdash \hat{I}_{ij}(a_j)$ and $\Lambda_i(g_i) \vdash \mathrm{B}_i^+(\mathrm{Nash}_i(a_i \mid a_j))$. Hence $\Lambda_i(g^i) \vdash \bigvee_{x_j} \left(\hat{I}_{ij}(x_j) \bigwedge \mathrm{B}_i^+(\mathrm{Nash}_i(x_i \mid x_j))\right)$. Thus $\Lambda_i(g^i) \vdash \hat{I}_{ii}(a_i)$.

Suppose $\Lambda_i(g^i) \vdash \bigvee_{x_j} \left(\hat{I}_{ij}(x_j) \bigwedge \mathrm{B}_i^+(\mathrm{Nash}_i(a_i \mid x_j))\right)$. We apply the belief-elimination operator ε to the both sides, we have, by (6),

$$g_i, \hat{g}_j \vdash_{\mathbf{0}} \bigvee_{x_j} (\mathrm{Nash}_j(x_j) \bigwedge \mathrm{Nash}_i(a_i \mid x_j)) .$$

This is the formal counterpart of the left-hand side of (2-1). ■

Proof of Theorem 6.C. We prove only (2-3). Suppose that a_j is a dominant strategy for $\hat{g}_j$. Let a_i be a best response to this a_j. Hence it follows from Theorem 6.B that $\Lambda_i(g^i) \vdash I_{ii}(a_i)$. Hence $\Lambda_i(g^i) \vdash \bigvee_{x_i} I_{ii}(x_i)$. Hence $\Lambda_i(g^i) \vdash \mathrm{B}_i^+(\bigvee_{x_i} I_{ii}(x_i))$.

Conversely, let $\Lambda_i(g^i) \vdash \mathrm{B}_i^+(\bigvee_{x_i} I_{ii}(x_i))$. Then it follows from Theorem 6.A that $\Lambda_i(g^i) \vdash \mathrm{B}_i^+(\bigvee_{x_i} \hat{I}_{ii}(x_i))$. Applying the belief-elimination operator ε to the both sides, we obtain $g_i, \hat{g}_j \vdash_{\mathbf{0}} \bigvee_{x_j} \mathrm{Nash}_j(x_j)$ by (6). Since the right-hand side does not depend upon g_i, we have $\hat{g}_j \vdash_{\mathbf{0}} \bigvee_{x_j} \mathrm{Nash}_j(x_j)$, using (4), which is the formal counterpart of the left-hand side of (2-3). ■

6.3 False Beliefs on the Other Player's Mind: From the Inductive Point of View

Consider a game $g = (g_1, g_2)$ where neither player has a dominant strategy. According to Section 5, each player i finds that he does not have a decision with respect to interaction structure $\mathcal{J}$ with $J_i = \{i\}$. Then it may be the case that players 1 and 2 try to make decisions with $\mathcal{J}^3 = (\{1,2\}, \{2\})$ and $\mathcal{J}^2 = (\{1\}, \{1,2\})$, respectively. Then the arguments of Subsections 6.1 and 6.2 are applied to each player. If we apply the arguments to both players simultaneously, then their subjective thoughts are different from the

objective reality and look "inconsistent". Now we show that this possibility could occur actually in our logical system and the resulting decisions are compatible with the inductively stable stationary state – Nash equilibrium – in the objective sense.

Recall that $g^1 = (g_1, \hat{g}_2)$ and $g^2 = (\hat{g}_1, g_2)$ are the vectors of payoff functions believed by 1 and 2. For these g^1 and g^2, we denote the unions of the sets of the following lists, respectively, by $\Lambda_1(g^1)$ and $\Lambda_1(g^2)$:

$$\mathsf{B}^1(\mathcal{J}^3, \mathrm{I(1\text{–}3)}), \mathsf{B}^1(\mathcal{J}^3, \mathrm{WD}), \mathsf{B}^1(\mathcal{J}^3, g^1);$$

$$\mathsf{B}^2(\mathcal{J}^2, \mathrm{I(1\text{–}3)}), \mathsf{B}^2(\mathcal{J}^2, \mathrm{WD}), \mathsf{B}^2(\mathcal{J}^2, g^2).$$

Now we can state the following theorem.

Theorem 6.D (Inductive Pitfall II): Suppose that each g_i allows no dominant strategies, but that (a_1^*, a_2^*) is an inductively stable stationary state. Suppose that each g_i satisfies condition SN. Then there are payoff functions $\hat{g}_1$ and $\hat{g}_2$ such that

(1): each $\hat{g}_i$ allows a dominant strategy;

(2): $g^i = (g_i; \hat{g}_j)$ satisfies condition Conc_i for $i, j = 1, 2$ $(i \neq j)$;

(3): players 1 and 2 have interaction structures $\mathcal{J}^3 = (\{1,2\}, \{2\})$ and $\mathcal{J}^2 = (\{1\}, \{1,2\})$, respectively;

(4): for $i = 1, 2$, $\Lambda_i(g^i) \vdash I_{ik}(a_k^*)$ and $\Lambda_i(g^i) \vdash \neg I_{ik}(a_k)$ for all $a_k \neq a_k^*$ and $k = 1, 2$;

(5): $\Lambda_1(g^1) \cup \Lambda_2(g^2)$ is consistent in $\mathrm{KD4}^2$.

Theorem 6.D can be proved in the same manner as the proof of Theorem 5.D, and we omit it. Here we give some examples of $g^1 = (g_1, \hat{g}_2)$ and $g^2 = (\hat{g}_1, g_2)$ in the game of Table 1.3:

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$
$\mathbf{s}_{11}$	$(5, \hat{0})$	$(1, \hat{2})$	$(3, \hat{0})$
$\mathbf{s}_{12}$	$(6, \hat{0})$	$(3, \hat{3})^*$	$(0, \hat{0})$

Table 6.1

	$\mathbf{s}_{21}$	$\mathbf{s}_{22}$	$\mathbf{s}_{23}$
$\mathbf{s}_{11}$	$(\hat{1}, 5)$	$(\hat{1}, 2)$	$(\hat{1}, 3)$
$\mathbf{s}_{12}$	$(\hat{3}, 1)$	$(\hat{3}, 3)^*$	$(\hat{3}, 2)$

Table 6.2

Tables 6.1 and 6.2 give $g^1 = (g_1, \hat{g}_2)$ and $g^2 = (\hat{g}_1, g_2)$. Each player i makes up his belief $\hat{g}_j$ on the other player's payoff function so that the observed action a_j^* has the unique dominant strategy for $\hat{g}_j$.

This theorem is similar to Theorem 5.D: the falsity here is caused by the lack of information about the other player's payoff function, while the

falsity stated in Theorem 5.D is due to the lack of the experiences of payoff values for each player himself.

As in Theorem 5.D, neither player cannot think about consistency, in which sense, this theorem states that the players may get stuck in the pitfall of induction. However, each player i may become conscious of the pitfall since he may deductively notice that some payoff functions for j different from $\hat{g}_j$ satisfy the same requirements except (5). Once each player becomes conscious about this pitfall, it would be the way out to communicate to each other about their payoff functions.

It is also possible that player 1 has interaction structure $\mathcal{J}^3 = (\{1,2\},\{2\})$ and 2 has $\mathcal{J}^1 = (\{1\},\{2\})$. In this case, their ways of thoughts are "consistent" but the contents of the thoughts may be still different. In this case, we can formulate this as a theorem parallel to Theorem 6.D.

References

[1] Honderich, T. (1995), Ed. *The Oxford Companion to Philosophy*, Oxford University Press. Oxford.

[2] Kaneko, M., (1997a), Epistemic Considerations of Decision Making in Games, IPPS. DP. 724. To appear in *Mathematical Social Sciences*.

[3] Kaneko, M., (1997b), Decision Making in Partially Interactive Games I: Game Theoretical Development, IPPS. DP. 743.

[4] Kaneko, M., (1998), Decision Making in Partially Interactive Games II: Game Logic Development. To be completed.

[5] Kaneko, M. and A. Matsui, (1997), Inductive Game Theory: Discrimination and Prejudices, IPPS. DP. 711, University of Tsukuba.

[6] Kaneko, M. and T. Nagashima, (1991), Final Decisions, Nash Equilibrium and Solvability in Games with the Common Knowledge of Logical Abilities, *Mathematical Social Sciences* 22, 229-255.

[7] Kaneko, M., and T. Nagashima, (1996), Game Logic and its Applications I. *Studia Logica* 57, 325-354.

[8] Kaneko, M. and T. Nagashima, (1997a), Game Logic and its Applications II. *Studia Logica* 58, 273-303.

[9] Kaneko, M., and T. Nagashima, (1997b), Axiomatic Indefinability of Common Knowledge in Finitary Logics, *Epistemic Logic and the Theory of Games and Decision,* eds. M. Bacharach, L. A. Gerard-Varet, P. Mongin and H. Shin, Kluwer Academic Press, 69-93.

[10] Moulin, H., (1982), *Game Theory for the Social Sciences*, New York University Press, New York.

[11] Nash, J. F., (1951), Noncooperative Games, *Annals of Mathematics* 54, 286-295.